2013

Addressing the Issue of Meta-Analysis Multiplicity in Education and Psychology

Joshua R. Polanin
Loyola University Chicago

LOYOLA UNIVERSITY CHICAGO

ADDRESSING THE ISSUE OF META-ANALYSIS MULTIPLICITY IN EDUCATION AND PSYCHOLOGY

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE GRADUATE SCHOOL

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

PROGRAM IN RESEARCH METHODOLOGY

BY

JOSHUA R. POLANIN

CHICAGO, IL

MAY 2013

ACKOWLEDGEMENTS

Few people in life stand in your corner no matter the circumstances or consequences. My parents, Rick and Terry, have supported my personal and professional goals without waver and I am forever indebted for their love and care. Moreover, they taught me the value of hard work, passion, and education. I would not have written this without their help.

I am also indebted to my advisor, mentor, and friend, Terri Pigott. My tenure at Loyola began by Dr. Pigott giving me an open opportunity that others may not have provided and it was her foresight and guidance that produced the academic I am today. Her input was especially apparent during many phases of this project. Indeed, this project's results far exceeded my expectations and that is a credit to her abilities as a mentor and teacher.

My family deserves much of the credit as well. My brother and sister, Brad and Krista, continue to endure an older brother who will never let them complain about term papers and title pages alone. My brother-in-law and his wife, Keigan and Kate, continue to teach me the joys of new family members. Similarly, my mother-in-law, Rhonda, endlessly provides support and love. Many, if not all, of my extended family have also supported me in various ways throughout this entire process as well.

Other individuals deserve mention for their continued support, most of whom I cannot list here. Alison Pigott spent a few weeks of her summer break reading abstracts

from journals she probably didn't know existed. Dan Kissel conspired to enjoy graduate school with me more than I thought possible. Ryan Williams continues to teach me about our field while providing a much-needed friend in the process. Dr. Dorothy Espelage readily provides sound wisdom and a kind heart; her passion for research pushes me to be a better methodologist. I must also thank my dissertation readers, Drs. Meng-Jia Wu-Bohannon and Steve Brown, for their support and guidance as well.

Finally, I am without words, feeling, or thought in absence of my wife, Megan. Since we met in class five years ago, she has been a sounding board, a guiding light, and a friend. Megan pulls me up from the depths and provides a gentle touch when I get too high. She stands with me in this great accomplishment and I can think of no treasure great enough to show my gratitude for her love and support. Her music engenders my continued grace and, to me, her presence is a song.

To Megan

To do or not to do a test of significance- that is a question that divides men of good will and sound competence.

-Robert F. Winch & Donald T. Campbell, The American Sociologist, 1969

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The concept of multiplicity, conducting multiple statistical significance tests in one study, has pervaded primary research over the last 7 decades (Hochberg & Tamhane, 1987; Keselman, Cribbie, & Holland, 1999; Neyman & Pearson, 1929; Tukey, 1949). This continued discussion was due to the fact that multiplicity increases the probability of committing a Type 1 error (i.e., deriving a false conclusion). Statisticians and methodologists, therefore, created methods to control this phenomenon, which in turn increases the validity of research results (Fisher, 1935; Keselman, Miller, & Holland, 2011). Little attention has been paid, unfortunately, to multiplicity in meta-analysis (Tendal, Nuesch, Higgins, Juni, & Gotzsche, 2011) and calls have been made for meta-analysis methodologists to address this critical issue (Sutton & Higgins, 2008). As such, the purpose and significance of this project was to answer these calls by formally quantifying the multiplicity of statistical test in meta-analyses published within education and psychology literature, and to ameliorate the problem of multiplicity errors through the advancement of Type 1 error corrections for meta-analyses.

To accomplish this goal, two methodological procedures were conducted. First, this author screened all citations in Psychological Bulletin and Review of Educational Research. From the citations that met inclusion criteria, 130 articles were randomly selected to code. The results revealed an alarmingly high number of statistical tests used per study ($\mu = 70.82$, $\sigma = 94.2$, $M = 46.5$). A major contributor to the number of

statistical tests utilized was the number of independent syntheses; the average study conducted 12.72 independent syntheses ($\sigma = 21.26$, $M = 5.0$). A multiple regression model predicting the number of statistical tests used per study found that the date of publication, number of studies included in the review, and the number of independent syntheses per review all were linear predictors.

For the second phase of the project, this author purposively selected four studies to investigate the potential use of Type 1 error corrections in meta-analysis. A formal guide to grouping the family of significance tests preceded the investigation given the plethora of opportunities to conduct statistical corrections. Given an appropriate guide, referred to as the "Timeline of Statistical Significance Testing", this author utilized four statistical correction techniques. The results provided by the review authors were compared to the results using the statistical corrections. Using the statistical corrections, an average of 3.33 conclusions would need to be modified per review.

The results of this project indicated a community of researchers becoming more reliant on statistical significance testing while simultaneously ignoring the consequences of multiplicity. It is no longer feasible to contend that meta-analysis is immune to Type 1 errors because of the use of the effect size. Meta-analysis methodologists must insist on clearer and directive research questions, protocols, and study parameters in addition to the consideration of multiplicity corrections. Failure to prevent further reliance on statistical significance testing in meta-analysis has the potential to prorogate the progress of cumulative science.

CHAPTER ONE

INTRODUCTION

The concept of multiplicity, conducting multiple tests of statistical significance within one study, has received much attention in primary research over the last 7 decades (Hochberg & Tamhane, 1987; Keselman, Cribbie, & Holland, 1999; Neyman & Pearson, 1928; J.W. Tukey, 1949) because conclusions derived from multiple tests have an increased probability of falsely rejecting a true null hypothesis. Researchers and statisticians, therefore, developed methods to control the probability of making such false conclusions. Although these procedures provided primary researchers a platform to reduce the probability of experiment- or family-wide error rates, little discussion has been conducted on multiplicity in meta-analysis (Tendal, Nüesch, Higgins, Jüni, & Gøtzsche, 2011). Multiple calls have been made for meta-analysis methodologists to address this important issue (Bender et al., 2008; Sutton & Higgins, 2008), yet few scientific advances have been put forth since Hedges and Olkin's (1985) suggestions. As such, the purpose and significance of this project is to answer these calls by formally quantifying the prevalence of multiplicity in meta-analyses within education and psychology and attempting to ameliorate the problem of multiplicity errors (i.e., Type 1 errors) through the advancement of methodological policy and statistical corrections for meta-analyses.

This is an important task because authors often laud meta-analysis as the answer to inconclusive results in primary research (Cooper, Hedges, & Valentine, 2009; Hedges

& Olkin, 1985; Hunter & Schmidt, 2004). The reason meta-analysis fosters clarity derives from the fact that meta-analysis relies on effect sizes synthesis. Effect sizes quantify the magnitude of the relationship between two variables and therefore are not susceptible to sample size bias (Kelley & Preacher, 2012). Null hypothesis testing, on the other hand, relies in part on the sample size from which the statistic is estimated in addition to the magnitude of the relationship (Cohen, 1994; Nickerson, 2000). In other words, given a large enough sample size, any relationship magnitude has the potential to be deemed statistically significant.

As such, examining the effect size magnitude offers an alternative to null hypothesis significance testing to form conclusions. A systematic review of the literature, given appropriate procedures, represents all known studies conducted on a particular topic of interest. The resulting average effect size, therefore, provides a more precise representation of the magnitude of the effect in question relative to primary research. Answers to challenging research questions, often with disparate primary study results, can be estimated using the procedures of quantitative synthesis.

The precision and scope of meta-analytic results engendered the use of the technique across most research disciplines (Cooper, Hedges, Valentine, 2009). Prior to meta-analysis, primary research was often unincorporated and disparate (Lipsey, 2007). Systematic reviews and meta-analysis synthesize large databases of information and maintain order within heavily researched enterprises. The inherent qualities of systematic review and meta-analysis have led to the increased usage of meta-analysis. Indeed, meta-analysis is now the standard of practice in fields like medicine (J. P. T. Higgins & Green,

2011), criminology (Lipsey & Wilson, 2001), and social work (Littell et al., 2005). These disciplines rely on the information provided by research syntheses to make decisions that affect both policy and practice.

In psychology and education, the use of meta-analysis as a research tool has increased exponentially over the past 25 years. Willams (2012) showed that the rate of published meta-analyses has increased steadily every year since 1990. In 2010 alone there were more than 800 meta-analyses published in the database *PsycInfo*. Education remained no different: During that same year, the database ERIC warehoused over 200 meta-analyses (Educational Resources Information Center, 2013). Clearly the technique has gained considerable market share of the researcher's conscience.

Given the prolific dissemination and increased usage of meta-analyses in decision-making for policy and practice, it is paramount to investigate and ensure the validity of reviews' results. Matt and Cook (2009) hypothesized that the validity of meta-analytic results resembled those of primary research. Following the traditions of Cook and Campbell (1979), the authors detailed three similar validity paradigms and their application to meta-analysis. The "threats to inferences about the causal nature of an association between treatment and outcome classes" discussed threats inherent to internal validity, specifically addressing studies with successful random assignment and primary study attrition (pg. 549). The second paradigm addressed threats to generalized inferences and included threats such as "rater drift" and "misspecification of causal mediating relationships" (pg. 550). Many of these issues determine whether a review's results represent valid conclusion. Although each validity paradigm contributes to the

conclusions derived from a review and ultimately the review's usefulness, a full review of each validity paradigm is outside the scope of this review.

The final threat, relevant to this project, is the "threat to inferences about the existence of an association between treatment and outcome classes" (pg. 540). Specifically, the authors briefly discussed the threat of "capitalizing on chance in meta-analysis" (pg. 544). The authors summarized the extent of the problem: "Although research syntheses may combine findings from hundreds of studies and thousands of respondents, they are not immune to inflated type 1 error when many statistical tests are conducted without adequate control for error rate" (pg. 545). The authors articulated that meta-analysis is not immune to the problem of Type 1 error. Moreover, the lay understanding of meta-analysis may be to reason that multiplicity errors do not occur, or rather occur less frequently, because of the large number of participants and studies. This assumption is without merit: Multiplicity errors occur given enough tests of statistical significance, regardless of study or participant sample size.

This issue has not been fully overlooked by the meta-analysis community. Bender et al. (2008) commented on the complexity of reviews and hypothesized that the complexity of research questions and divergent syntheses resulted in multiplicity errors due to multiple comparisons. Although the authors identified some of the reasons for multiplicity, the authors failed to quantify the extent of the problem. On the other hand, Cafri, Kromrey, and Brannick (2010) sought to quantify the extent of Type 1 errors and corresponding power through a systematic review of meta-analyses in psychology. The results indicated that only 14% of statistical tests had power less than .80 (pg. 252), but

up to 78% of the studies' conclusions could be the product of Type 1 errors. Unfortunately, the authors failed to explicate how to reduce Type 1 errors or the reasons for multiplicity.

Although previous attempts have been made to deal with the problem of multiplicity in meta-analysis, the investigations remain in infancy. Moreover, meta-analysis, as a research tool, remains in the developmental stages as well. In order to bolster the validity of meta-analysis' results and further the understanding of the problem, this project sought to fill in the gaps left by other methodologists as well as expand the awareness of the multiplicity problem. The following sections detail the basics of meta-analysis and null hypothesis testing as well as ways to combat the rate of Type 1 errors.

Research Questions

Given the influence statistical significance testing has on meta-analysis, further research is warranted on the precision and prevalence of statistical tests used to bolster meta-analysis claims. This project will seek to quantify this information using the following research questions as a guide:

1. What is the prevalence of statistical significance testing in published meta-analyses within education and psychology?

    a. How many hypothesis tests are conducted per study?

    b. How many independent syntheses are conducted per study?

    c. What statistical adjustments for multiplicity (e.g., Bonferroni), if any, are utilized?

    d. Are there moderators that capture the differences in study-level rates?

2. How should methodological guidelines improve to reflect multiplicity issues in meta-analysis?

a. How would one correct for multiple tests of statistical significance?

b. Do the conclusions from current meta-analyses change when multiplicity corrections are applied?

To answer these questions, this author conducted a review of meta-analyses. This author reviewed and coded 130 peer-reviewed, published meta-analyses across two social science review journals: *Psychological Bulletin* and *Review of Educational Research*. The first question was addressed through the quantification of multiplicity: Each meta-analysis was coded for the various types of statistical significance testing and independent syntheses. Additional study-level information, such as type of multiplicity correction, was also collected. A multiple regression model was used to predict the number of statistical significance tests. This author answered the second question by purposively selecting 4 of the 130 articles and applying current methods of statistical adjustments not usually utilized in meta-analyses currently. These adjusted statistical results were compared to published results. Implications for future meta-analysts were considered.

Significance of the Study

Little discussion persists with regard to multiplicity in meta-analysis and few guidelines exist with regard to how to handle the treat of spurious findings. The methods utilized today often derived during meta-analysis' beginning stages (i.e., Hedges & Olkin, 1985) or have failed to gain popularity (Laird et al., 2005). Bender et al. (2008)

recently advocated for increased evidence and guidance from the meta-analysis methods community. Understanding the prevalence of statistical tests and how they impact the validity and robustness of research synthesis results have the potential to engender solutions to this problem.

Furthering the methods of meta-analysis is an important pursuit because quantitative reviews have recently gained popularity among practitioners and policy-makers (Field, 2003). Indeed, meta-analysis is the preferred method of data-based decisions across many domains including medicine (J. P. T. Higgins & Green, 2011), psychology (Cooper, 2010), education (Pigott, 2012), and criminology (Lipsey & Wilson, 2001). The conclusions derived from meta-analyses inform policy and practice and thus it is imperative that the findings are valid.

CHAPTER TWO

LITERATURE REVIEW

Null Hypothesis Significance Testing

The discussion surrounding null hypothesis significance testing derives from primary research (NHST;Cohen, 1994; Howell, 2006). Although the estimations and test statistics vary widely, the central theme has remained mostly unchanged. A researcher wishes to test empirically a hypothesis using sample data. Most often, the null hypothesis is represented by H_0 and the alternative by H_1. The null hypothesis is generally conceived as the null distribution being equal to the distribution in question. Depending on the research question and statistical test, the analyst determines the alpha level and test direction (i.e., one or two-tailed). The analyst retains the null hypothesis when the statistical test fails to provide enough evidence to support the alternative hypothesis. In contrast, if enough evidence is produced (i.e., the test statistic is greater than a predetermined critical value), then the analyst rejects the null hypothesis in favor of the alternative hypothesis. This framework holds for parametric or nonparametric assumptions, continuous or discrete scales, and primary or meta-analytic data analysis.

Error Rates

The error rate is defined as the probability of falsely rejecting a "true" null hypothesis (Howell, 2006); this type of error is commonly referred to as a Type 1 error

(Agresti, 2009; Nickerson, 2000). A Type 2 error occurs when an analyst falsely accepts the null hypothesis, when in fact the null hypothesis is false. It follows, then, that the analyst can correctly retain or reject the null hypothesis thus drawing the correct conclusions (Table 1).

Table 1. Type of Errors

	Null Hypothesis is True	Null Hypothesis is False
Reject Null Hypothesis	Type 1 error (i.e., False positive	Correct conclusion (i.e., True positive)
Fail to Reject Null Hypothesis	Correct conclusion (i.e., True negative	Type 2 error (i.e., False negative)

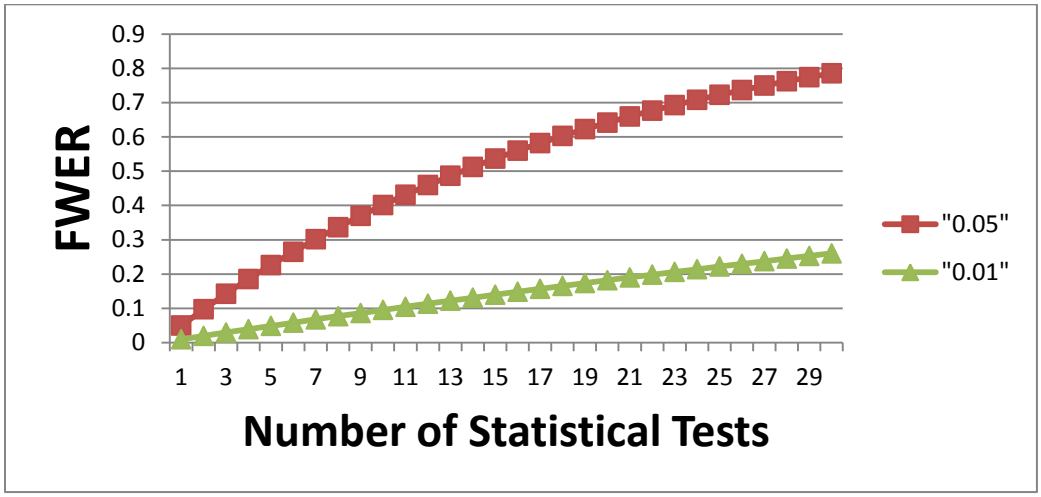
The error rate for a single statistical test is commonly referred to as the comparison-wise error rate (CER) or alpha (α) (Bender & Lange, 2001). For example, if the resulting p-value from a direct statistical test is $p = .03$, then 3% of the time the z-statistic calculated will be this large completely from chance and, therefore, the rejection of the null hypothesis false. It is essentially the probability the reviewer is willing to accept of making the incorrect conclusion and, as such, an arbitrary number (Cohen, 1994).

When the reviewer conducts a number of statistical tests within a study, the resulting error rate becomes a "familywise" error rate (Tukey, 1949). It has been shown repeatedly in the literature that increasing the number of statistical tests will increase the probability of committing a Type 1 error (Keselman, Miller, & Holland, 2011; Neyman & Pearson, 1928; J.W. Tukey, 1949).This can be shown as the following:

$$\text{CER } \alpha = \alpha' \qquad \text{(EQ 1)}$$
$$\text{Familywise error rate (FWER)} = 1 - (1 - \alpha')^c$$

where α' is the analysts' chosen error rate for a given comparison (e.g., α' = .05), and c is the number of comparisons. For instance, an experiment with merely two comparisons will increase the familywise error rate, or probability of having at least one Type 1 error, $1 - (1 - .05)^2 = .0925$. Increasing the number of comparisons to 3 or 4 increases the probability of having a Type 1 error exponentially (.143 & .186, respectively). A study with 100 comparisons almost certainly will falsely reject at least one true null hypothesis (FW = 99.4%). Decreasing the alpha leads to a corresponding decrease in the probability of Type 1 error, but it does not eliminate it entirely (Figure 1). The probability of committing a Type 1 error increases each time a test of statistical significance is conducted.

Figure 1. Family-wise Error Rate



It should be noted that the above equation holds only for independent tests of statistical significance. Dependent tests of statistical significance require a slightly modified FWER calculation that incorporates the covariance (or correlation) of the statistical tests. Because the correlation is rarely known or given, it is difficult to examine

the corrected familywise error rate empirically. Bender and Lange (2001) noted that for most hypothesis testing the assumption of independence is viable. Further, dependence is a relative term and some debate exists as to what constitutes dependent tests. Some authors have posited corrections for dependent statistical tests, but the most common tests assume independence.

A Type 2 error, it should also be noted, consists of a different premise. The researcher is said to commit a Type 2 error when a false null hypothesis is wrongly accepted. In other words, the researcher should reject the null hypothesis but incorrectly determines the null hypothesis true. With the avocation of many multiple significance tests in the biological sciences, a recent movement is afoot to decrease the rate of Type 2 errors as well as Type 1 errors.

Meta-Analytic Approach

Meta-analysis is not immune to family-wise error rates because it utilizes tests of statistical significance (Borenstein, Hedges, Higgins, & Rothstein, 2009). Each synthesis conducts a series of statistical tests that answers slightly different questions. The application of current meta-analysis methods requires the reviewer to conduct null hypothesis testing while often disregarding the probability of false results inherent in multiple null hypothesis testing. A number of issues must be addressed to understand fully the issues surrounding meta-analysis multiplicity. The following sections outline the current systematic review and meta-analysis best practices, the reasons for multiplicity inherent in meta-analysis, and the ways to control for the effects of multiple tests.

Conducting a Meta-Analysis

The process of conducting a meta-analysis consists of a series of research steps. Although many systems and guidelines pervade the literature (Cooper, 2010; Littell, Corcoran, & Pillai, 2008), all follow similar procedures. To be clear, conducting a meta-analysis constitutes synthesizing quantitative effect size estimates from multiple primary studies. To collect the effect sizes, a systematic review process is conducted. The difference being that meta-analysis constitutes the synthesis of effect sizes to calculate an average effect while systematic review is the collection of primary studies to form a general opinion.

The first step, like all scientific research, is to devise research questions from a scientific problem. This process requires a deep intellectual knowledge of the research problem and an intricate understanding of the primary literature. The reason for an intricate understanding of the literature is twofold. First, by nature, meta-analysis seeks to combine multiple primary research articles. If only one (or very few) articles on a given topic exist, a meta-analysis will fail to provide helpful information. Second, the reviewer must devise a list of variables and specify outcomes of interest. Although this protocol often will receive iterative changes throughout the course of a typical meta-analysis, the more detailed the protocol at the first stage the smoother the research process. Both of these steps require knowledge about the area under study.

The second stage of a typical meta-analysis is the literature search and retrieval. At this stage, the researcher focuses on a comprehensive and systematic review of literature databases (Littell, et al., 2008). Multiple iterations of parameter fields, search

terms, and Boolean characters are required of the researcher to locate all possible primary research articles. Often, the researcher must contact the articles' authors in hopes of locating further articles and would be wise to consult the bibliographies of retrieved articles. A recent movement also has advocated for the searching of "fugitive literature", or literature that is not commonly read or disseminated. Primary research presented at conferences, conducted for thesis or dissertation purposes, or unpublished manuscripts constitute this type of literature. Only after complete saturation of literature will the researcher abate the literature search.

Information extraction constitutes the third stage of meta-analysis. The researcher, after retrieving all pertinent primary articles, must code all relevant study information. Wilson (2009) advocated for a comprehensive coding process that enables the researcher myriad options post-extraction. To this end, the meta-analyst, again, must have a fundamental understanding of the literature; however, regardless of insight, this process, too, is iterative. Cooper (2010) advocated for a process where the researcher creates a coding tool then extracts information from 1-2 primary studies. The researcher then critiques the coding form before proceeding to code the remaining studies. Orwin and Vevea (2009) argued that the researcher would do well to code each article twice with multiple independent raters. Often, an inter-rater reliability will be reported to confirm the extracted information's validity.

During the third stage, the meta-analyst must also extract summary statistics to estimate an effect size. The effect size, as mentioned previously, is the magnitude of the relationship between two variables (Borenstein, 2009). The magnitude of the effect size is

not directly impacted by sample size; however, sample size impacts the sampling error associated with the effect size and must be accounted for in the eventual synthesis. Further, the process of standardization is what allows meta-analysts to synthesize multiple study effect sizes; without the standardization process one would have little to meta-analyze.

Effect size estimates derive from various data configurations. The three most common effect sizes remain the standardized mean-difference, odds ratio, and correlation coefficient. The transcription of the effect size calculations is outside the scope of this review, however, as they have been described elsewhere in detail (see Hedges & Olkin, 1985 for a review). The variance calculations also differ across the effect sizes but again remain well-discussed in the literature (Lipsey & Wilson, 2001).

The fourth stage in conducting a meta-analysis is combining the effect sizes from each study. Hedges and Olkin (1985) found that weighting each study by the inverse of its sample variance was an efficient way to represent the overall average. Using this technique, larger studies receive greater influence because larger studies, relative to small studies, are more precise. The sampling variance is calculated differently for each effect size (see Cooper, 2010; Lipsey & Wilson, 2001). It should be noted that the weights differ slightly under fixed and random effects models. A fixed effect model assumes that the only reason effect sizes differ is because of sampling variation. Random effects models, on the other hand, consider both sampling variation and between-study variation associated with study characteristics. Relative to the fixed effect model, larger studies do not receive equally greater weight under a random effects model.

Estimating an overall weighted average effect size is not the final process. Indeed the meta-analyst must test for the presence of heterogeneity among the effect sizes and the presence of effect size moderators. Important substantive questions may be answered by exploring how the effect size varies across subgroups and study characteristics. In addition, Cooper (2010) suggested presenting each study's effect size and summary statistics in a summary table. Meta-analysts must also describe policy implications and generalizability in this final step.

Statistical Significance Tests Conducted in Meta-Analysis

One aspect removed from the previous discussion was the use of statistical significance testing in meta-analysis. A number of statistical significance tests coincide with each quantitative synthesis phase (Borenstein et al., 2009) and each determines whether the estimated statistic was due to sampling error alone. Each of these tests follows the traditional null hypothesis testing framework of primary study design. In theory, a meta-analyst is interested in testing whether some statistic derives from a null distribution. The meta-analyst proceeds by testing the null hypothesis that the null distribution is equal to the estimated distribution. From these two distribution, the meta-analyst may calculate the probability (p-value) that the null distribution is equal to the estimated distribution. Given a small enough probability (traditionally $p < .05$), the meta-analyst rejects the null hypothesis, stating instead that the estimated distribution is significantly different from the null distribution. The distributions of test statistics differ; an analyst must estimate a different distribution for each test statistic.

In practice, however, the meta-analyst needs only to calculate a test statistic and compare that statistic against a critical value to determine statistical significance. For most reviews, the first set of test statistics calculated are tests of the overall average effect. Assuming a fixed effect model, generally, the significance test conducted utilizes a z-test, and can be written:

$$z = \frac{|\overline{ES}|}{SE_{\overline{ES}}} \quad \text{(EQ 2)}$$

where $|\overline{ES}|$ represents the absolute value of the average overall effect size and $SE_{\overline{ES}}$ is the standard error of the overall average effect size. The standard error is represented by:

$$SE_{\overline{ES}} = \sqrt{\frac{1}{\sum w_i}} \quad \text{(EQ 3)}$$

where w_i is study i's weight, either from the fixed or random effects models. The resulting z-statistic is distributed as a standard normal variate with critical value of 1.96 representing a two-tailed significance of $p < .05$ (Lipsey & Wilson, 2001).

The next procedure is then to test the overall distribution of effect sizes for homogeneity. Again assume that the analyst chose a fixed effect model. The first hypothesis test will measure the amount of variability between studies and decide if this is greater than expected by sampling error (or within-study variance). Formally, an overall Q statistic is calculated by:

$$Q = \sum_{i=1}^{k} W_i \, (Y_i - M)^2 \quad \text{(EQ 4)}$$

where W_i is the weight of the study i, Y_i is the mean effect size (any metric) of study i, and M is the overall average effect size. To test whether significant variation exists, the Q statistic is compared to a chi-square distribution with *df* = k-1.

Given significant variation in the distribution of effect sizes, the meta-analyst has a number of choices to explore said variability: 1) dependent subgroup analyses, 2) categorical moderator analyses using the *Q*-Between statistics, 3) pairwise comparisons, and 4) meta-regression models. Each type of analysis uses a test statistic to determine quantitatively whether the effect could have occurred by chance alone. Given a test statistic greater than a pre-determined critical value, the meta-analyst again rejects the null hypothesis.

The meta-analyst may choose to explore variability using non-independent (i.e., *dependent*) subgroup analyses. Essentially the effect sizes are grouped according to study-level characteristics. A word of description is required to delineate the difference between *independent* synthesis and *dependent* subgroup analyses. Independent syntheses do not share common effect sizes. An effect size derived within one study is not used to calculate more than one average effect size. For instance, a meta-analyst may choose to split effect sizes by the type of outcome and then synthesize each different outcome independently. In contrast, a meta-analyst may choose to group effect sizes into different subgroups using the same effect size across the groups. These subgroups, therefore, are inherently dependent because they share a common effect size. Whether independent or dependent, however, Equation 2 is again utilized to test the overall average effect size.

Given that significant variation exists, the analyst conducts subgroup analyses. These analyses, generally, observe study-level differences. This is analogous to primary studies estimating group differences, for example an intervention versus a control group. The goal is to calculate a test statistic that provides evidence against the null hypothesis (H_0). Analyses with only two groups test for a simple difference, analyses with more than two groups observe differences from the grand mean.

Under a fixed effect model, three methods exist to test study-level differences; however, because most analyses use only two of the three methods, the first two prominent methods will be discussed. Borenstein et al.'s (2009) first method compares only two groups using the traditional z-test. The analyst first calculates the overall average effect size (M_A & M_B) and within-group variances (V_{M_A} & V_{M_B}) for each group. A confidence interval and z-test for each group is permissible at this stage, but unimportant to calculate the z-test of the differences. Next, the analyst calculates the difference between the overall average effect sizes of each group:

$$Diff = M_A - M_B \quad \text{(EQ 5)}$$

and the standard error:

$$SE_{Diff} = \sqrt{V_{M_A} + V_{M_B}} \quad \text{(EQ 6)}$$

The z-test of the differences is then:

$$Z_{Diff} = \frac{Diff}{SE_{Diff}} \quad \text{(EQ 7)}$$

This z_{diff} follows a standard normal distribution where a non-adjusted, two-tailed test of significance at the .05 level yields a critical value of 1.96.

Another option is available to the analyst where the number of groups is greater than or equal to 2. Borenstein et al. (2009) described this method as analogous to a one-way ANOVA model because the variance portioned into within- and between-groups and tested against a chi-square distribution. To conduct the ANOVA-like model, a *Q* statistic is calculated for each group (e.g., Q_A, Q_B, Q_C). Then, the sum of the Q "within" is taken, formally:

$$Q_{Within} = \sum_{j=1}^{p} Q_j \quad \text{(EQ 8)}$$

where *p* is the number of groups, and Q_j is the Q statistic for group j. Using the Q statistic calculated in eq. XX, the $Q_{Between}$ is calculated:

$$Q_{Between} = Q_{Total} - Q_{Within} \quad \text{(EQ 9)}$$

The $Q_{Between}$ is then compared against a chi-square distribution with *df* = p - 1 to test the omnibus hypothesis of group differences (H_0: $M_A = M_B = M_C$). The analyst can also test the homogeneity of Q_{Within} or Q_j. The Q_{Within} is the average variance within groups while each Q_j is a measure of the magnitude of the variation within group j. Each is compared against a chi-square distribution, with *df* = k (number of studies) – p and *df* = k (number of studies in group p) – 1, respectively. Finally, Q_{Total} is tested to ensure overall variance; here, the chi-square distribution has *df* = k – 1. Of course, if variability persists within each group, and the overall test of $Q_{Between}$ is significant, it is warranted to test further moderators.

To sum, a number a direct statistical tests of significance are conducted using the ANOVA-like model. The Q_{Total} has one test, the $Q_{Between}$ and Q_{Within} two more, and then each group has its own direct test, in this example 3 more. It is not uncommon to present the p-values for each of the statistical tests.

Finally, a meta-analyst may choose to utilize a univariate or multiple regression model. The former simply regresses one explanatory variable on the dependent variable, in this case, the effect size from study *i*. Higgins and Thompson (2004) suggested using this method when the number of studies to be synthesized is less than 10. The test of interest is the test of the slope, H_0: $\beta = 0$. Traditionally, the meta-regression test is based on the Z-distribution, which is formally written as:

$$Z = \frac{\beta}{SE_\beta} \quad \text{(EQ 10)}$$

where β is the slope of the explanatory variable regressed on the dependent variable and SE_β is the standard error of the slope estimate.

A variation of the univariate model is the multiple predictor regression model. This model produces an overall Q-Total statistic as well as various Z statistics for each slope in the model. The slope tests of statistical significance follow EQ 10. The Q-Total is simply the weighted sum of squares, reflecting the "total dispersion of studies about the grand mean" (Borenstein et al., 2009; pg. 208). The test statistic follows a chi-square distribution with df = k – 1. The model tests for the presence of at least one significant explanatory variable in the model.

Multiplicity in Meta-Analysis

Despite the laudation meta-analysis receives for the use of effect sizes (Cohen, 1997), the process of conducting meta-analysis leads to the use of statistical significance

testing (EQs. 1-10). Often the conclusions reached by these significance tests leads to further significance testing. For example, Borenstein et al. (2009) cautioned readers not to conduct moderator analyses given a non-significant overall Q test of homogeneity. Similarly, pair-wise comparisons should not be invoked given a non-significant Q-Between significance moderator test (Hedges & Olkin, 1985).

Therefore it is clear that meta-analyses utilize tests of statistical significance. The rate at which a meta-analysis utilizes tests of statistical significance, however, requires some explanation. Methodologists have implicitly hypothesized three reasons why a meta-analysis may conduct multiple significance tests.

Explaining Variation among Effect Sizes

One reason why reviews carry out tests of statistical significance is to explain variation among the effect sizes, usually in the form of subgroup or moderator tests. As explained previously, a subgroup analysis generally tests for the overall average effect size within one dependent group. Bornstein's (1989) review provided an example of this type of moderator analysis. The review first estimated an overall average effect size of the relationship between exposure to stimulus and affect. The review then grouped the effect sizes based on the type of stimuli, maximum number of stimulus presentations, duration of exposure, type of measure, delay, and average age. The same effect was represented across multiple subgroups, but the author conducted a test of overall average effect for each subgroup. The author conducted 31 z-tests of statistical significance, one for each subgroup.

A moderator analysis, in contrast, groups effect sizes into independent levels and conducts another overall test of homogeneity. In this framework, the analyst also conducts tests of Q-Within for each level as well as pair-wise comparisons. Connell and Goodman (2002) synthesized effect sizes that observed children's internalizing and externalizing behavior problem as a function of their parent's mental health problems. An average correlation, as well as corresponding overall tests of average effect and homogeneity, was estimated for each type of problem behavior and for each parent. For each of the four independent syntheses, the authors conducted a series of one-way ANOVA models. A total of eight categorical variables were hypothesized to moderate the heterogeneity in the outcomes, each with varying numbers of levels. The number of categorical moderators tested varied across each of the four synthesis, but a total of 22 tests of Q-Between were conducted as well as 106 tests of Q-Within. However, the review did not report conducting pair-wise comparisons within the levels. Including the overall tests and eventual meta-regression models, this study conducted 170 tests of statistical significance.

A similar form of moderator test, meta-regression, uses a regression framework to test the linear relationship of the moderator and the effect sizes. Unlike primary research, it is acceptable to conduct multiple univariate regression models as well as simultaneous multiple regression models (Borenstein et al., 2009). Either model type, however, uses tests of statistical significance. For example, Bus (1995) synthesized correlations between parental book reading and three outcome measures, reading achievement, emergent literacy, and language skills as well as a composite variable. For each outcome, including

the composite variable, the authors conducted a series of univariate regression models testing the relationship between seven moderators and the effect sizes. A total of 28 meta-regression tests of statistical significance were conducted, and a total of 36 tests across the entirety of the study.

Bender et al.'s (2008) framework for splitting effect sizes into independent groups is one explanation. Since each synthesis generally conducts, at a minimum, two tests of statistical significance (i.e., Overall test of average effect size, Overall test of homogeneity), it follows then that reviews with more syntheses will conduct a greater number of tests of statistical significance.

Splitting Effect Sizes into Multiple Syntheses

The traditional meta-analytic procedures described above implicitly assumes that the effect sizes to be synthesized represent a single underlying, *a priori* construct. The meta-analyst, for example, is interested in synthesizing the effectiveness of cultural awareness training on students' empathy attitudes. The synthesis collects all appropriate primary studies, each with one intervention and control group, measuring a form of empathy, at one time point. Theoretically, ambiguity and uncertainty should remain low because the meta-analyst only collects an effect size when an empathy measure is present. For the meta-analyst, this is the best-case scenario.

Yet, conducting a meta-analysis is rarely this straightforward (Cooper, 2010). Primary studies measure slightly different, but related outcomes; participants vary widely; observations range from immediate post-test to 12-month follow-up. The permutations are endless and unregulated. For the meta-analyst, the task of deciding

which effect size to synthesize and which to throw out is daunting. Each decision could bias the final estimate.

To combat this challenge, meta-analysts have two primary options: Lump or Split effect sizes. Lumping represents a meta-analysis that synthesizes all related outcomes. If the author can make a reasonable case of relation, then the effect size is included in the review. Splitting represents the opposite; the effect sizes are grouped into meaningful subsets, each representing a slightly different construct or population of interest (Weir, Grimshaw, Mayhew, & Fergusson, 2012).

The results of splitting is a large number of *independent* syntheses, or multiple average effect sizes where only one effect size from each study can be represented. What often occurs when conducting a review is that unexpected or previously unknown groups of effect sizes may be found. If this is the case, the meta-analysts must decide whether to split these effect sizes into separate groups or lump all effect sizes, regardless of population type or time point, into one larger group. For example, Byrnes, Miller, and Schafer (1999) synthesized gender differences in risk taking where risk taking was represented by a number of different behaviors. Only one effect size from each study was grouped into one of the behaviors. Each of the 16 different independent syntheses differed by the type of primary study outcome.

In contrast, Abrami and colleagues (2008) synthesized intervention effect sizes to increase critical thinking skills. Regardless of the type of critical thinking skill, one overall average effect was calculated. The resulting lumped average effect size was then

subdivided into subgroups, but these subgroups were *dependent* because the same effect size could be represented in multiple levels.

When reviewing a published meta-analysis, the reader's only source of information about whether the meta-analyst lumped or split is effect sizes the review's analysis. If the review authors split up the effect sizes into separate independent studies, it is determined that the review split effect sizes. If one overall average effect is estimated, then the authors lumped all effect sizes.

In actuality, the decision to lump or split effect sizes may occur at numerous occasions (Goodyear-Smith, van Driel, Arroll, & Del Mar, 2012). The first decision point occurs when a meta-analyst is determining the research questions for the scope of the review. Will the review ascertain all effect sizes related to a particular outcome of interest or limit the inclusion to only very specific types of studies? Review authors again must make this same decision when explicating the inclusion/exclusion criteria. Will certain studies be excluded given particular parameters of the population or will all populations be included?

Even after detailed research questions and inclusion/exclusion criteria, unknown effect sizes of interest may persist. In the data extraction phase, when effect sizes are calculated, the review author must again decide whether each individual effect size or one composite effect size will be extracted. Finally, during data analysis, the review authors must again decide whether to split the effect sizes into independent syntheses or lump all effect sizes into one large overall effect size. It is the culmination of all these decisions that is rendered in the final publication.

When a reviewer determines that effect sizes must be split into multiple independent syntheses, there are a theoretically infinite number of reasons. Bender et al. (2008) hypothesized that there are most likely four primary reasons to "split" the effect sizes into independent syntheses. The following section will describe the possible scenarios where the author might decide to split the synthesis, conducting theoretically-independent syntheses within the same study. A decision tree is also illustrated to elucidate when and where the splits are likely to occur (Figure 1).

Multiple Outcomes. The goal of meta-analysis is to synthesize primary study outcomes of interest. Although in theory the research question limits the researchers' options with regard to the outcome of interest, the meta-analyst is often at the mercy of the primary study researchers. For instance, a study on the effects of a drug for depression may seek to improve the quality of life, negative or positive affect, behavioral interactions, or a host of other important characteristics all of which may be considered an important outcome.

The problem for meta-analysts is twofold. First, the meta-analyst must attempt to decide *a priori* which outcomes are of interest. This can be a difficult task if one is unfamiliar or unaware of the myriad outcomes. The second problem is how to handle the multiple outcomes. Often a theoretical argument can be posited simply to synthesize all of the outcomes as they measure the same "construct" (Cook & Campbell, 1979). Rationalizing in this manner ameliorates the need to conduct multiple outcome syntheses. When this is theoretically or practically impossible, the meta-analyst must decide to conduct multiple syntheses in the same study.

Multiple Effect Sizes. Not unlike a scenario with differing outcomes, primary studies may elect to measure outcomes in varying metrics. Fortunately, myriad calculations exist to transform a primary study's effect size to that of a common effect size (Lipsey & Wilson, 2001; Wilson, 2009). For example, a review may contain studies that estimate standardized mean-differences and odds ratios effect sizes. The meta-analyst will often elect to transform the odds ratios to standardized mean-differences in order to synthesize all outcomes in a similar metric.

Of course, the problem arises when the effect sizes differ greatly or are incomparable. A common problem, especially for novice reviewers, is the difference between two-group independent standardized mean-differences and a single-group standardized-mean difference gain score. The former estimates the difference between two independent groups, usually a treatment and control, while the later estimates the differences at two time points but with one group. Although both statistics produce an estimate of mean-difference, the statistics represent vastly different meanings. Moreover, a transformation will never be available to estimate either type.

Therefore, the meta-analyst must decide, again preferably *a priori*, how to handle this and similar scenarios. The reviewer may decide to eliminate all effect sizes (or studies) that fail to meet the specified types of effect sizes required. This is a common technique in the methods literature (Cooper, 2010; Cooper et al., 2009). The other option, again, is to include the studies but conduct multiple syntheses within a single meta-analysis. This option will allow for a greater understanding of the literature and bolster results, but will introduce multiplicity of statistical significance tests.

Multiple Groups. Rarely will two studies replicate the same methods; rather, an iterative approach will be taken that compares different types of intervention groups with controls. A factorial design with increasing treatment dosage is one example where this scenario may occur within one study (Campbell, 1957; Campbell, Stanley, & Gage, 1963).

The job of the meta-analyst, then, is to decide the most appropriate effect size to extract. One option is to simply ignore group comparisons made by primary studies that fail to meet an *a priori* protocol decisions. The problem with this technique, however, remains that this could bias the results and subsequent conclusions. Often the meta-analyst will choose, therefore, to extract effect sizes for each group separately. For example, a review may be interested in combining the effectiveness of an educational program to reduce test anxiety. One group of studies may implement an intervention aimed at increasing only the student's parental involvement, while another group of studies introduces a program to include both a parental and teacher component. The meta-analysis, in this example, will synthesize the groups independently. As these two scenarios constitute diffuse interventions, the meta-analyst may decide to synthesize the effect sizes independently for each scenario.

Multiple Time Points. A primary study interested in the effects of an intervention rarely requires information about its effectiveness immediately after the completion of the program. Rather, researchers who implement interventions investigate the program's effectiveness given a certain duration after its completion. As such, the

researcher will collect observations both post-intervention and at some designated follow-up time. An effect size can be calculated for each time point.

Bender et al. (2008) posited that two distinct problems arise from multiple time point observations. The first problem is that not all primary studies measure follow-up at the same time, if at all. Some studies measure follow-up at 1 month, others 10 months, and others 10 years. It is difficult to rectify how best to handle each of these seemingly disparate effect sizes. The second problem occurs when an effect size is included for a project that has yet to conclude. For example, a planned intervention collects multiple observations across the duration of the program, not necessarily post-intervention. In order to publish findings as soon as possible, the researchers publish intervention effects while the program is still being conducted. This effect size estimate may constitute a different construct from those collected post-intervention.

The meta-analyst must decide, *a priori*, how to handle studies that measure outcomes at different time points. It is sometimes possible to synthesize studies that collect data at similar time points. For instance, a reviewer may decide to synthesize studies that utilize any length of treatment durations, essentially lumping all effect sizes. More likely, the analyst conducts multiple syntheses within a single meta-analysis. Here, the meta-analysis synthesizes outcomes, for example, at both post-treatment and at follow-up.

Lumped vs. Splitting Effect Sizes Example. Consider a scenario such as the one represented in Figure 2. The first column represents the four reasons Bender et al. (2008) hypothesized. The second column represents the lumped effect size scenarios while the

third column represents a review where the effect sizes are split. At each level, a decision is made to either lump or split the effect sizes. Given a lumped strategy at each level, the review synthesizes one group of effect sizes. This procedure therefore renders only two tests of statistical significance, one for the overall tests of average effect and the overall tests of effect size homogeneity. Of course tests may be conducted to explain effect size heterogeneity, but these are not considered in this example.

Figure 2. Lumped vs. Splitting Effect Sizes Example

Level	Lumped Synthesis	Split Syntheses
Outcome	Depression & Anxiety	Depression or Anxiety
Effect Size	One-Group & Two-Group Designs	One-Group or Two-Group Designs
Treatment	Cognitive- Behavioral Therapy & Acceptance and Commitment	Cognitive-Behavioral Therapy or Acceptance and Commitment
Time Point	Post-Treatment & Follow-up	Post-Treatment or Follow-up
Total θ	1	16
Total Tests of Significance*	2	32

Notes: *Total Tests of Significance includes only the overall tests of the average effect size and overall tests of effect size homogeneity.

The second example represents a worse-case scenario for splitting effect sizes. At each level, the meta-analyst determines that the effect sizes should not be synthesized together and therefore splits the effect sizes into separate groups. Doing so at each level constitutes multiplying the number of effect sizes by a factor of 2 at each step. For example, one synthesis might represent the depression outcome, following a two-group design, using cognitive-behavioral therapy, measuring the outcome at post-treatment. A

second synthesis might represent anxiety outcome, following a two-group design, using cognitive-behavioral therapy, measuring the outcome at post-treatment. As a result, the review synthesizes 16 separate syntheses for a minimum of 32 tests of statistical significance.

Updating a Review

The final reason multiple tests of statistical significance can occur is due to review updating. When a meta-analysis is conducted, it is inevitable that further primary research will be conducted after its publication (Trikalinos & Ioannidis, 2005). Indeed, methodological research has indicated that effect sizes change dramatically after the completion of early studies, and smaller effects often require a greater length of time to publication (i.e., time-lag bias). Therefore, the authors advocated for the continual update of meta-analyses. To meet this requirement, a meta-analyst would recalculate the overall average effect size (and subsequent subgroup/moderator analyses) after each newly conducted trial. Regardless of update schedule, recalculating the overall average effect size to update the literature constitutes multiplicity of statistical tests.

The difference between this type of multiplicity and the other previous forms is that updating a review consists of conducting multiple tests of statistical significance among multiple studies. The previous considerations of multiplicity assumed that all tests of statistical significance occurred within a single review. To include review updates as part-in-parcel as multiplicity in a single review requires one to reconsider the parameters of the family-wise error rate. A liberal understanding of the family-wise error rate would consider review updates as multiplicity, given that the same (or very similar) population

of effect sizes is utilized to conduct the significance testing. However, the spirit of family-wise error rate calculations assumes that the family of significance testing occurs within a single study. This project, therefore, will henceforth only consider multiplicity when it occurs within a given review, and not consider review updates at present.

Controlling Type 1 Error

The classical view of controlling Type 1 error attempts to restrict false conclusions within a given family of tests. Therefore, it is essential to support with theory, *a priori*, what constitutes the family of tests. The most common type of family occurs under the analysis procedures of ANOVA (in primary research). Here, the analyst's goal is to determine if one or more of the groups departs from the grand mean. Given departure, the analyst continues to estimate individual group differences. Myriad other scenarios arise similarly that introduce multiple tests of significance.

Regardless of how the family is defined, to control familywise error rate, the analyst ensures that:

$$\text{FWER} \leq \alpha \qquad \text{(EQ 11)}$$

for all possible hypotheses tested for a given family (Lehmann & Romano, 2005). The principle here remains that the total FWER should total equal to or less than the analyst's specified alpha. Again, selection of alpha is completely arbitrary, but $\alpha = .05$ or $\alpha = .01$ are the common critical values in the literature.

Procedures vary with regard to the type of error rates one wishes to control, and multiple types exist (Hochberg & Tamhane, 1987). Weak control of FWER occurs when the researcher decreases the probability of Type 1 error for some but not all hypotheses.

Pena, Habiger, and Wu (2011) explained that weak control transpires when at least one true null hypothesis is falsely rejected, given that all null hypotheses were in fact false. This is rare and often undesirable. In contrast, strong control of FWER transpires when the researcher decreases the probability of Type 1 error for all situations. Strong control of type 1 errors allows for any combination of false and true null hypotheses. Methods of controlling the error rate, it follows, correspond to the level of control desired by the analyst. Generally, a balance between weak and strong controls is warranted. Newer methods that control Type 1 error seek such a balance.

Classical Control of Type 1 Error

Neyman and Pearson (1928) receive credit for the first discussion of the Type 1 error rate. It was Fisher (1935), however, that introduced the two classical procedures for control of Type 1 error and the two approaches divide how most researchers classify and utilize adjustments (Hochberg & Tamhane, 1987). Myriad adjustments persist with regard to methodological design, the number of groups, and statistical components such as the variance or dependence. The following section merely summarizes and describes the common classical procedures.

The first types of adjustments are generally referred to as “single-step” approaches. The defining characteristic, as the name implies, remains that the researcher only introduces one adjustment procedure prior to a decision. Again, the analyst must determine the level of alpha, either at the comparison-wise or family-wise level. Most often, the researcher will choose an alpha level at or near .05.

Fisher's (1935) famous Bonferroni procedure pervades all literature still today. This procedure is simple because the researcher merely divides the desired FWE alpha by the number of tests conducted. More formally, this is written as:

$$\alpha^* = \frac{\alpha'}{c} \quad \text{(EQ 12)}$$

where α* represents the new rejection p-value for each comparison, α' is the analysts pre-determined alpha, and c is the number of comparisons to be made. For example, a researcher may be interested in hypothesis testing the difference between multiple treatment groups and calls for 3 comparisons. Using Fisher's correction, if α' = .05, then $\frac{.05}{3}$ =.017. Each hypothesis is then tested against p = .017. It can be easily seen that the FWER is equal to .05 and therefore controls Type 1 errors at this level for the entire family of tests.

The Bonferroni procedure unfortunately becomes intrusively restrictive rapidly. Recent advances in computer technology and research design have engendered expansive families where it is not uncommon, especially in the biological sciences, to test hundreds of hypotheses simultaneously. Indeed, a family size of 15, suggested as a plausible upper level in psychology (Kaselman et al., 2011), creates conservative alpha levels ($\alpha^* = \frac{.05}{15}$ = .003). The simplicity of the procedure, however, maintains its utilization with researchers.

This ultra-conservative alpha rate, however, inspired researchers to modify the popular technique (Aickin & Gensler, 1996). Stepwise procedures utilize the structure of the hypotheses and test each in order of their p-values. Hochberg and Tamhane (1987)

explained the procedure consists of the researcher testing one hypothesis and making a decision to reject. If the tested hypothesis fails to provide enough evidence against the null hypothesis (i.e., the null hypothesis is retained) then testing ceases. Generally, the alpha level changes with each test. So-called step-down or step-up procedures differ by how the procedures order the hypotheses. A step-down procedure orders the hypotheses with the largest p-value in descending order, and testing starts with the largest p-value. A step-up procedure orders the hypotheses with the smallest first in ascending order; the test starts, then, with the smallest p-value. As with all previous examples, it is important to remember that only families of tests are grouped together.

Holm's (1979) sequentially rejective multiple test procedure combines the Bonferroni theorem with a step-up procedure. First, the analyst orders the p-values in ascending order, starting with the smallest, $p_{(0)} \leq p_{(1)} \leq p_{(2)} \ldots p_{(k)}$, which correspond to $H^{(0)}$, $H^{(1)}$, $H^{(2)} \ldots H^{(k)}$ hypotheses. Second, the analyst chooses an alpha, where α is $0 \leq \alpha \leq 1$, usually $P(1\text{-}.95)$. Third, the procedure starts by testing the smallest p-value against the Bonferroni procedure (EQ. 12). Fourth, if the first hypothesis is retained then all hypothesis testing stops and all hypotheses are retained. The procedure continues, however, if the null hypothesis is rejected. Fifth, given a null hypothesis rejection, a new alpha level is devised, formally given by:

$$\alpha^* = \frac{\alpha}{n-k} \qquad \text{(EQ 13)}$$

where α* is the new alpha level, α is the researcher's specified FWER (e.g., 05), n is the number of tests to be made, and k is the ordered p-value number. For example, the

second alpha ($\alpha^*_{(2)}$) corresponding to $p_{(2)}$ is $\frac{.05}{n-2}$. Sixth, the procedure continues until a null hypothesis is retained. At this point, all other hypotheses are retained.

For example, a researcher may be interested in testing 6 hypotheses. The corresponding vector of p-values is [.004, .009, .018, .045, .12, .96]. The first hypothesis alpha level therefore is $\alpha^*_{(0)} = \frac{.05}{6-0}$ = .008. Therefore the first null hypothesis is rejected and the procedure continues. The second hypothesis alpha level is $\alpha^*_{(1)} = \frac{.05}{6-1}$ = .01 and the second null hypothesis is also rejected. The third hypothesis is then compared to $\alpha^*_{(2)} = \frac{.05}{6-2}$ = .0125 and is therefore retained. Subsequently, all other hypotheses are also retained.

The procedure is more powerful relative to the Bonferroni procedure. Under the Bonferroni procedure, only 1 null hypothesis is rejected; under the assumptions of this procedure the number of null hypothesis rejections is two. However Holm's (1979) procedure is more conservative relative to no correction. The analyst who chooses not to utilize a correction procedure, in this scenario, rejects 4 hypotheses. Hochberg and Tamhane (1987) confirmed this anecdotal evidence through a series of simulation studies.

Recent Advances Controlling Type 1 Error

The classic procedures utilized an assumption where the number of hypotheses tested remained small and, usually, known *a priori*. The inherent problem is that it is often of import to test many multiples of hypotheses. Indeed, primary research and meta-analysis increasingly estimate more tests because including multiples of tests, instead of conducting research one experiment at a time, is cost-effective and efficient. Fisher

(1926) agreed that benefits exist to conducting multiples of hypotheses (instead of one at a time), "If we ask her [nature] a single question, she will often refuse to answer until some other topic has been discussed" (from Hochberg & Tamhane, 1987, p. 5).

Lehmann and Ramano (2005) understood that the restrictiveness of traditional corrections and the researchers need to test many multiple hypotheses were two opposing forces. Moreover, the conservatism of traditional corrections had created a completely new problem: Type 2 errors. The concern shifted from a risk of rejecting too many null hypotheses to rejecting too few. To combat this problem, the authors rationalized that it is of interest to the researcher to accept the possibility of making "k" (usually 2 or 3) false rejections to guard against many Type 2 errors. The authors termed the phrase "k-FWER" to allow for the possibility of the k false rejections. When k = 1, the FWER returns to the traditional error rate of one false rejections. Considering the above discussion of weak versus strong controls for Type 1 error, this new procedure is considered a weak control because it allows for the possibility of Type 1 errors.

Allowing for this possibility, however, initiated a new wave of procedures. Keselman et al. (2011) summarized the promising theories in a recent paper. The first, Lehmann and Romano's (2005) generalized Holm procedure, derives from the Holm (1979) procedure discussed earlier. The researcher again orders the p-values, starting with the smallest in ascending order: $p_{(1)} \ldots \leq p_{(k)} \ldots \leq p_{(m)}$ (note the slight change in notation) where the p-values correspond to $H_{(1)} \ldots H_{(k)} \ldots H_{(m)}$ hypothesis tests. Following Keselman et al., the procedure proceeds as follows:

Step 0. Let $i = 1$, k and α are chosen by the experimenter.

Step 1. If $i \leq k$, go to Step 2. If $k < i \leq m$, go to Step 3. Otherwise stop and reject all of the hypotheses.

Step 2. If $p_{(i)} > \frac{k*\alpha}{m}$, go to Step 4. Otherwise, set $i = i + 1$ and go back to Step 1.

Step 3. If $p_{(i)} > \frac{k*\alpha}{m+k-1}$, go to Step 4. Otherwise, set $i = i + 1$ and go to Step 1.

Step 4. Reject $H_{(m)}$ for $m < i$ and accept $H_{(m)}$ for $m \geq i$.

Returning to the example presented in the above section, suppose a researcher conducted $H_{(6)}$ hypothesis tests, and [.004, .009, .018, .045, .12, .96] are the $p_{(m)}$ vector of p-values. Assume α = (1-.95) or .05 and k = 2. Step 1 directs the researcher to proceed to step 2 because $i < k$ (1 < 2). The next step calculates the new alpha rate; since $p_{(1)} < \frac{2*.05}{6}$ < .017, the researcher sets $i = 1 + 1$ and returns to Step 1. Again, $i \leq k$ (2 ≤ 2), so the researcher returns to Step 2. Again, $p_{(2)} < \frac{2*.05}{6} < .017$, so the researcher sets $i = 2 + 1$ and returns to Step 1. This time, k < i ≤ m (2 < 3 ≤ 6) so the researcher proceeds to Step 3. At step 3, $p_{(3)} > \frac{2*.05}{6+3-1} > .0125$, so the researcher moves to Step 4. Here, the researcher rejects the first two hypotheses and retains the remaining four.

Lehmann and Romano (2005) also adapted a step-up procedure from Hochberg (1988) design. The major difference between the two is the order of the p-values. Contrary to the generalized Holm procedure, the generalized Hochberg procedure orders the p-values from the largest to smallest in descending order, $p_{(1)} \ldots \geq p_{(k)} \ldots \geq \ldots p_{(m)}$. Following the procedures implemented by Keselman et al. (2011), the researcher utilizes a series of steps:

Step 0. Let i = m, k and α are chosen by the experimenter, where m is the number of comparisons and k is the number of true null hypothesis rejections.

Step 1. If $i > k$, go to Step 2. If $1 \leq i \leq k$, go to Step 3. Otherwise stop and *accept* all of the hypotheses.

Step 2. If $p_{(i)} \leq \frac{k*\alpha}{m+k-i}$ go to Step 4. Otherwise, set $i = i - 1$ and go back to Step 1.

Step 3. If $p_{(i)} \leq \frac{k*\alpha}{m}$, go to Step 4. Otherwise, set $i = i - 1$ and go to Step 1.

Step 4. Reject $H_{(m)}$ for $m < i$ and accept $H_{(m)}$ for $m \geq i$.

Using the above example, an analyst calculates the unadjusted p-values from six comparisons [.004, .009, .018, .045, .12, .96]. The first step directs the researcher to proceed to step 2 because 6 > 2. At the second step, .96 > $\frac{2*.05}{6+2-6}$ > .05, so the researcher returns to the first step. Again, the researcher is directed back to the first step because .12 > $\frac{2*.05}{6+2-5}$ > .033. The next p-value is compared against $\frac{2*.05}{6+2-4}$ = .025 which is again greater than $p_{(3)}$. The $p_{(4)}$ has an adjusted alpha of $\frac{2*.05}{6+2-3}$ = .02. Here, the $p_{(4)}$ is less than the adjusted alpha (.018 < .02) and therefore the researcher is directed to step 4. Because $p_{(4)}$ is less than the adjusted alpha, all further hypotheses and the current are rejected. Thus, this procedure resulted in 3 null hypotheses rejected.

Keselman et al. (2011) advocated for the step-up procedure. The authors provided two reasons. One, the step-up procedure will always produce at least as many rejections as the step-down procedures because of their natures. Second, the step-up procedure, however, is more powerful relative to step-down procedures. The reason is that the critical values associated with each adjustment remain smaller and thus less evidence is

required to reject. The procedure will provide protection against Type 1 errors, but is less conservative than its counterpart or predecessors.

Controlling Error Rates in Meta-Analysis Using Statistical Corrections

Given the multiplicity of correction literature available for the primary researcher, it is reasonable to assume that corrections are available to the meta-analyst as well. The fact remains, however, that controlling for error rates in meta-analysis lacks sound methodological guidance. Indeed, Bender et al. (2008) and Tendal et al. (2011) called for greater awareness and application of even the simplest primary research corrections. Borenstein et al. (2009) summarized the sentiment, stating that "there is no consensus that conducting many comparisons can pose a problem, [and] there is no consensus about how this problem should be handled. As will be shown below, the community of meta-analysis methodologists advocates for only a few techniques.

Most of the limited literature on corrections in meta-analysis derived from Hedges and Olkin's (1985) original work. The authors advocated for two well-known techniques. The first was simply to correct using a modified Bonferroni technique in the context of subgroup analyses. The modified Bonferroni, also known as Dunn's (1961) correction, is formally written:

$$\alpha^* = \frac{\alpha}{2 * c} \qquad \text{(EQ 14)}$$

where α^* is the researcher's new alpha level, α is the researcher's alpha level for the family of comparisons, and c is the number of comparisons. For example, consider a family of tests that estimate differences based on treatment location, and the researcher plans to conduct 5 study-level comparisons. The researcher simply divides the alpha

(usually .05) by two times the number of comparisons, $p^* = \frac{.05}{2 * 5} = .005$. The procedure is nearly identical to the primary study procedure and therefore encounters similar problems (i.e., too conservative with many comparisons). Nevertheless, it is strong protection against true null hypothesis rejection.

The second method advocated by Hedges and Olkin (1985) was Scheffe's S (1959) procedure. This technique generally is reserved for planned comparisons, but has been generalized for post hoc comparisons as well. The meta-analyst compares the squared z-statistic (eq. 10) against the chi-square distribution with "c'" (number of comparisons - 1) degrees of freedom. In the previous example, 5 comparisons were made. A chi-square distribution with 5 *df* has a critical value of 11.05; thus the squared z-statistic must exceed 11.05 to reject the null hypothesis.

Hedges and Olkin (1985) briefly compared the two procedures as well. The authors maintained that the Bonferroni procedure was "quite powerful" when the number of comparisons was small (p. 161). On the other hand, when the number of comparisons increased, the Scheffe procedure produced more precise rejection estimates. The authors failed to mention, however, what constituted a "small" or "large" amount. Therefore it is the meta-analyst practitioner to decide which procedure represents the appropriate correction.

It should also be mentioned that the authors proposed these corrections in the context of subgroup analyses and multiple comparisons. No correction (or mention) for other tests of statistical significance were provided. Granted, the above procedures could

easily be adapted to fit other statistical tests outside of subgroup analysis (e.g., multiple syntheses). It is nonetheless interesting that no mention of the problem exists.

Laird et al. (2005) has attempted to utilize a new technique of correction. In contrast to the previous method, Laird et al. advocated for a stepwise procedure entitled the "false discovery rate" method (FDR; Benjamini, 1995). The FDR is defined as "the expected value of V/R, where R is the number of rejected null hypotheses and V is the number of rejected null hypotheses that are true" (Keselman et al., 2011, p. 427). To utilize this procedure, the researcher again orders the p-values starting with the smallest in ascending order (i.e., $p_{(1)} \ldots \leq p_{(2)} \ldots \leq p_{(m)}$) that correspond to *m* null hypotheses. The researcher's comparison-wise alpha rate is written as:

$$\alpha^* = \frac{i}{m} * \alpha \qquad \text{(EQ 15)}$$

where *i* is the ordered p-value, *m* is the number of comparisons, and α is the familywise alpha (e.g., .05). The procedure begins by starting with the largest p-value and working backwards; the procedure stops when the first null hypothesis is rejected. If no hypotheses are rejected then all hypotheses are accepted. For example, returning to the above example, consider the ascending vector of p-values [.004, .009, .018, .045, .12, .96]. The largest p-value ($p_{(6)}$) is compared against $\alpha^* = \frac{6}{6} * .05 = .05$; thus, the largest hypothesis is retained and the procedure continues. The second and third largest p-values are compared against .042 and .033, respectively, and are also retained. The fourth largest p-value ($p_{(3)}$), however, is compared against $\alpha^* = \frac{3}{6} * .025$ and is rejected. Therefore, $H_{(1)}$, $H_{(2)}$, and $H_{(3)}$ null hypotheses are also rejected.

Laird et al. (2005) advocated for the FDR procedure in the context of a relatively new form of meta-analysis in genetics. This technique uses many multiples of statistical significance tests and thus traditional correction methods are often overly conservative. The effect sizes estimated, however, differ slightly from traditional effect sizes generated from social science research. The authors mentioned that the correction should be utilized in the greater context of meta-analysis. To date, no correction method has been proposed (to this author's knowledge) in the medical or social science meta-analysis methodological literature.

Higgins and Thompson (2004) addressed "spurious findings" in a meta-regression context (pg. 1663). The authors conducted various simulations to estimate the proclivity of Type 1 errors in meta-regression models and advocated for the use of a permutation test to combat these errors. The procedure followed a process that required the meta-regression models and the data be available in raw format rather than simply relying on ad hoc procedures that adjust utilize the p-values. While this is an important contribution to the literature of multiplicity corrections in the context of meta-regression, unfortunately the method could not be tested using p-values alone. The authors raised important concerns, however, with regard to Type 1 errors inherent in meta-regression models. As such, these models and their corresponding tests of statistical significance should not be excluded when considering multiplicity in meta-analysis and should therefore be included in the overall discussion.

Finally, Borenstein et al. (2009) suggested that the decrease to .01. The authors stated that increased usage of statistical tests, as well as an increase in perceived power,

provided ample leverage to suggest the decrease in alpha. The logic of decreasing the critical value to .01 follows that of Fisher (1935) in that he suggested both .05 and .01 as critical values (by estimating test statistic critical values for each level of alpha). Yet, Borenstein et al. failed to provide an explanation as to why the p-value of .01 was chosen. While it will guard against spurious findings, the arbitrary number lacks statistical justification. Nevertheless, it is one more option that review authors may turn.

Other Ways to Control Error Rates in Meta-Analysis

Other ways are available to control the rate of Type 1 error rates in meta-analysis instead of ad hoc multiplicity corrections. The common trait that these methods share is simply their lack of use among meta-analysts. If primary research's history can be a lesson for meta-analysis, the traditional use of statistical significance testing will reign supreme in meta-analysis for years to come. Nevertheless, there remain a number of options available to the interested meta-analyst.

One computationally intensive and theoretically imposing approach involves combining dependent effect sizes, such as related outcomes or time points, into one synthesis (Hedges, Tipton, & Johnson, 2010). The technique is different from simply taking the average of multiple related effect sizes within one study, instead using all available data in a multi-level framework. Thus, more information is actually collected and utilized and as a result more robust conclusions

There are two major advantages to this approach. The first is that the assumption of independence of effect sizes is maintained. When a synthesist derives multiple effect sizes from the same group of participants, regardless of whether the outcomes are

theoretically divergent, the effect sizes are inherently dependent. This approach mitigates this dependence. The second advantage to this approach is relevant to the current project. By synthesizing multiple dependent effect sizes simultaneously, say related outcomes, the meta-analysis consequently reduces the number of independent syntheses (or subgroup analyses) conducted within a given review. The consequences from a Type 1 error perspective, are obvious: A decrease in independent syntheses should, theoretically, reduce the number of statistical significance tests.

Another way to decrease the rate of statistical significance testing, indeed completely eliminate it, is to switch from a frequentist perspective to a Bayesian perspective (Sutton & Abrams, 2001). Compared to the traditional meta-analytic approach which uses a frequentist framework, the Bayesian framework provides model uncertainty instead of model probability (actually non-probability, *p*-values). From a practical perspective, the results of a Bayesian meta-analysis translate directly to "quantities of interest, for example, the probability that patients receiving drug A have a better median survival than B" (pg. 279). The use of Bayesian methods also necessitate that the analyst provide a "prior distribution" that, albeit arbitrary and susceptible to bias, requires the analyst to have a deep understanding of the problem.

Relevant to this project, the results of Bayesian meta-analyses are expressed in terms of model uncertainty (i.e., a "prediction interval") instead of the form of a significance tests (Welton, Sutton, Cooper, Abrams, & Ades, 2012). The Bayesian analyses utilize a Markov Chain Monte Carlo (MCMC) model to predict, to 95% certainty, the true value of the underlying parameter. Because these are model based

analyses, instead of frequentist approximations, the results provide a more precise answer. Moreover, the use of multiple models within a single study does not alarm the Bayesian because each model represents a different distribution; therefore, there is no threat of Type 1 error because the paradigm of Type 1 error does not exist in Bayesian analyses.

The problem, of course, with Bayesian analysis is twofold. One, the approach is highly computer intensive, requiring extensive background in MCMC modeling and computer software programming (Sutton & Abrams, 2001). Although this problem is quickly becoming a relic of the past, it is still a concern today. Second, the approach is simply under-utilized because Bayesian analyses have failed to gain popularity relative to frequentist approaches: Bayesian analysis remains difficult to perceive in the eyes of many frequentists (Berger, 2006). In order to utilize these techniques fully, the community of meta-analysts must first embrace the technique.

The final approach is perhaps the most radical but easiest to employ: Stop using statistical significance testing altogether. This radical paradigm has not yet been fully explored in meta-analysis, but the practice would have plenty of support from primary research methodologists and statisticians. In the book *What if there were no Significance Tests* (Harlow, Mulaik, & Steiger, 1997), myriad authors postulated a research world without the constraints of p-values or power analyses. A well-cited contribution came from the late Jacob Cohen, in a chapter (earlier reproduced as a journal article) entitled *The Earth is Round ($p < .05$)* (Cohen, 1994). In this chapter, Cohen described his personal and professional discomforts with significance testing. Most chief among them,

what he believed was wrong with null hypothesis significance testing: “It does not tell us what we want to know, and we so much want to know what we want to know that, out of desperation, we nevertheless believe that it does” (pg. 22). In other words, Cohen felt that the significance test failed to answer the most basic of questions, whether A differs from B, and therefore could not be of use (at least to the extent he felt it was being employed) to researchers and practitioners.

A more practical solution derived from a leader in meta-analysis. Schmidt and Hunter (1997) addressed the concerns of many p-value supporters by advocating for the usage of confidence intervals. They argued that little would be lost through the eradication of statistical significance testing while much would be gained simply through the predicted interval approach. Tukey (1991) offered a similar approach for those who feared that the loss of significance testing would render multiple comparisons useless.

Summary

Multiplicity of statistical testing is a problem for primary as well as meta-analytic research: The more tests of statistical significance conducted, the greater the probability of deriving false conclusions. Although meta-analysis, compared to primary research, attempts to utilize effect size synthesis as the primary method of summary, null hypothesis significance testing is relied upon throughout the analysis stages. The question remains, however, how often and to what extent has meta-analysis become reliant on statistical significance testing as a means of answering research questions? The primary purpose of this project, therefore, is to answer the call for greater clarity with regard to the use of statistical significance testing.

Of course, simply knowing how often tests of statistical significance are utilized is only one part of the problem. As shown previously, little clarity exists on how to control the rate of Type 1 errors in meta-analysis. Therefore a second purpose of this project was to ascertain how often meta-analytic results control for Type 1 error as well as to derive a set of methods for meta-analysts to utilize to control for future Type 1 errors. The following sections will detail the procedures to assess the usage of statistical significance testing in meta-analysis and one way to combat threats to statistical conclusion validity. The final section of this project will delineate the impact of these procedures and future directions of this line of research.

CHAPTER THREE

METHODS

This review of reviews, also known as a meta-review or overview, systematically selected published meta-analyses from two leading review journals. Two distinct phases, corresponding with the research questions, directed the methodology. This author first sought to screen and select review articles in order to quantify the use of statistical significance testing (i.e., multiplicity) and determine potential moderators of usage. The second phase answered the second research question by creating a process to guide the grouping of statistical tests for multiplicity corrections. Four reviews were then purposively selected to test the new guideline procedures in conjunction with existing, suggested multiplicity corrections. The corrected results were compared to the published results.

Phase I

Sample

Published meta-analyses served as the primary observational unit. Included reviews were published between 1986-2011 and derived from either Psychological Bulletin (PB) or the Review of Educational Research (RER). The date parameters were chosen because they represented the era when meta-analysis grew in popularity as well as established a primary set of recommendations and procedures. These journals were chosen purposively for a number of reasons. First, these journals produced reviews that

aligned with the project's inclusion criteria. The purpose of this project was to review education- and psychology-focused meta-analyses and each journal published reviews specific to these disciplines. Second, the journals offered high quality studies that each underwent an intensive peer-review process. This process, theoretically, should produce reviews that met the current meta-analysis standards of practice. The studies published were heralded as paramount research studies. Third, systematically reviewing two specific journals allows for replicability and conduction ease. Future meta-analysis methodologists, should they choose, may repeat this project's methods easily. Simply reviewing meta-analyses in their published form, organized by date, allowed this author to conduct the review with relative ease and efficiency.

In anticipation for the review of these two journals, this author investigated and estimated the prevalence of potentially included meta-analyses. This author first estimated the number of articles per issue by sampling two issues within each year and counting the number of articles. This process provided an average number of articles per issue, and used further to calculate the number of articles per year. Then, this author estimated the percentage of articles that met the inclusion criteria by screening the titles and abstracts of a randomly selected proportion of the issues. The resulting estimation was summed for each journal. The initial procedure estimated that that PB published 282 (95% CI: 280,284) potential reviews while RER had only 143 (95% CI: 141, 145) potential reviews. Again, these were only estimates to provide a framework for the selection process.

To ensure that the sample of studies represented the populations of studies (i.e., all of the potentially includable studies from the two journals), this project utilized a

stratified-random sampling technique. The date of publication and journal served as the two main strata. Published reviews from 1986-2000 or 2001-2011 represented the first two substratum; the second two substratum were represented by the two journals. As such, this sampling design constituted a 2x2 sampling frame. Proportional allocation (Lohr, 1999) was used to maintain the unbalanced samples across the strata. Within each of the four stratum, this author planned to randomly select 20% of the reviews. This process resulted in at least 95 total meta-analyses selected for final analysis.

Inclusion/Exclusion Criteria

The reviews must have met several criteria to be included in the sample (Appendix A). First, the reviews must have appeared in either PB or the RER from 1986-2011. Second, the reviews must have focused on either a psychological or an educational topic. Third, the reviews must have been a quantitative synthesis (i.e., meta-analysis) and not merely a systematic review. Studies that reviewed other meta-analyses (e.g., meta-reviews, umbrella reviews, etc.) were also eliminated. Fourth, the review must have utilized a technique that combines effect sizes to calculate point estimates and confidence intervals (or have the potential to calculate confidence intervals). This eliminated vote counting procedures or descriptive studies. Fifth, the studies must have presented the results of the synthesis in a quantitative manner.

The selection process followed a two-stage process. This author first read the title and abstract of each review. Each study was designated “include”, “unsure but include”, or “discard”. The review process followed a standardized format (see Appendix A). This process eliminated many of the articles. Articles labeled as “unsure” were downloaded

and screened using a standardized tool (Appendix B). Unlike the previous stage, each study was labeled “include” or “discard”.

All types of outcomes, effect sizes, and synthesis techniques were included. No preference was given to continuous or dichotomous outcomes because both types of studies are used widely across education and psychology. Reviews that synthesized bivariate correlations were also included because these studies were also susceptible to multiplicity. However, studies that failed to provide information about the number of effect sizes, the reasons for splitting the effect sizes into independent syntheses, or the number of moderator or meta-regression tests were excluded.

The remaining studies constituted the final sampling frame and were randomly selected in accordance with the stratified sampling design. In the event that a study was included but should have been excluded from the sample, this author replaced the study by randomly selecting another study from that study’s stratum to replace it.

Measurements

A codebook was constructed to extract study-specific information in a standardized manner (Appendix B). The goal of a review codebook is to extract information efficiently while also being inclusive enough so as not to warrant post-extraction reexaminations (Littell et al., 2008). As such, the codebook allows for variance between studies by providing broad directions for coders while maintaining a standard format. The resulting codes provide rich description of each study. The result of this project’s extraction process provided information that allowed statistical analysis of the issue of multiplicity.

Given these guidelines for a codebook, this author constructed a codebook that was efficient but inclusive. Four major sections guided the coding process. The first section coded basic information about the review: author names, number of authors, date of publication, the title of the article, the publication source, and if the review received funding. The second section detailed the study's meta-analytic characteristics; here, this coder detailed the topic, a qualitative description of the purpose, and type of study. Because both observational and intervention studies were included in the review, type of study was an important code. Also included in the section was a description of the purpose of the meta-analysis. This code was generated to inform differences within different types of meta-analyses (i.e., observational vs. experimental). Five primary purposes were hypothesized. The first two, differences and bivariate relationships, were considered observational meta-analyses. A difference meta-analysis observed gender or racial differences on a hypothesized outcome; a bivariate relationship meta-analysis synthesized correlations. Efficacy and effectiveness were two purposes that constituted experimental research. The purpose of an efficacy meta-analysis was to test a broad research question; for example, a meta-analyst might be interested in the efficacy of counseling on depression. An effectiveness review, on the other hand, directly targeted specific types of programs. In the case of depression, an effectiveness review might target psychodynamic programs specifically. The fifth meta-analysis purpose was deemed prediction. This specification has the possibility to cross both observational and experimental. Finally, a catch-all "other" code was also utilized.

Other important methodological characteristics of the meta-analyses were also coded. The codebook coded for whether the study included: a graphical plot, primary

study quality coding, a power analysis, grey literature, Cohen's effect size classifications, publication bias analyses, and model specifications. In addition, codes were included for whether the review was an update of a previous review and who the review author cited when referring to the meta-analytic details.

The second half of the codebook detailed the issues associated with multiplicity. This section extracted specific information on the number of statistical tests and syntheses. Each independent synthesis was coded separately. For instance, if the review authors split the effect sizes into two groups (i.e., by the outcome), then each synthesis was coded separately. For each synthesis, then, this author coded the number of: overall tests of average effect size, overall tests of homogeneity, overall tests of subgroup average effects, Q-Between tests, Q-Within tests, and meta-regression slope tests. In addition, the number of reported significant tests was also included. Coding each synthesis in this manner allowed for a calculation of the total number of independent syntheses as well as the total number of statistical tests per review.

The fourth section coded information about the review's attempts to correct for multiplicity. This section coded whether the review authors discussed the issue of multiplicity, if the review adjusted for multiplicity, the study's alpha level, and what technique (if any) were used to control Type 1 errors.

Analysis

The first analysis task was to describe the sample of studies descriptively. This included a description of the review's general characteristics, multiplicity characteristics, and other methodological aspects. The results were subset by the source of publication for greater description.

A second important task, considering the research questions, was to determine the extent to which statistical significance tests were utilized. To answer this question, this author calculated the use of each type of statistical significance test overall, by the source, and by the date of publication. In addition to the number of tests conducted by the type of test, a composite variable representing the total number of significance tests was also calculated. The equation below delineates how the summary statistic was calculated:

$$\begin{aligned}\text{Total Tests} = {} & \text{Number of Overall Z Tests} + \text{Number of Overall Q Tests} + \\ & \text{Number of Q-Between Tests} + \text{Number of Q-Within Tests} + \\ & \text{Number of Z-Tests for Moderator} + \text{Number of Meta-Regression Tests.}\end{aligned} \qquad \text{(EQ 16)}$$

To be clear, sensitivity analyses performed by the review authors were not included in this calculation. Sensitivity analyses were labeled by the review authors as simple exploratory hypothesis tests and were generally treated as auxiliary to the primary analyses. Howell (2006) recommended not to include sensitivity analyses when considering Type 1 errors. In addition, this total number does not factor in the number of significant tests. To explore how these tests interact within a given review, a correlation matrix of the type of tests was also constructed.

Another important task that this task this project answered was the extent to which reviews "split" or "lumped" effect sizes. Bender et al.'s (2008) recommendation for effect size splitting was utilized as a template. Because reviews failed to follow the hypothesis proposed by Bender et al, this author detailed the numerous other ways that reviews split the synthesized effect sizes. The analyses included in this section sought to understand the rate at which authors rationalized different split reasons. In addition, the analysis explored how the splitting of effect sizes impacted the rate of statistical

significance testing. Following previous analyses, the results are presented in a descriptive manner.

The final analysis sought to understand the relationships between review characteristics and the use of statistical significance testing. For this analysis, a multiple regression models was utilized. The assumptions of multiple regression were checked prior to conducting the analysis. Because the dependent variable, total tests of statistical significance, followed a slightly positive skew, the log-transformation was utilized. To follow the assumption of independence, all variables were aggregated to the study level. A correlation matrix was also created to estimate the impact of multicollinearity. The author hypothesized that the number of syntheses, the number of effect sizes included in the review, the date of publication, the observational type of review, and whether the review was funded would predict the number of tests conducted. A number of other variables were included as controls. The multiple regression model can be represented by:

$$Y_i = \beta_0 + \beta_1 * (\text{Source})_1 + \beta_2 * (\text{Number of Authors})_2 + \beta_3 * (\text{Date of Publication})_3 + \beta_4 * (\text{Funded})_4 + \beta_5 * (\text{Discuss Multiplicity})_5 + \beta_6 * (\text{Adjust Alphas})_6 + \beta_7 * (\text{Number of Studies})_7 + \beta_8 * (\text{Number of Split Reasons})_8 + \beta_9 * (\text{Independent Syntheses})_9 + \beta_{10} * (\text{Experimental}) + \beta_{11} * (\text{Both})_{11} + \beta_{12} * (\text{Full Update})_{12} + \beta_{13} * (\text{Partial Update})_{13} + r_i \quad \text{(EQ 17)}$$

where Y_i represented the log-transformed total number of statistical tests for study *i*, β_0 represented the constant, β_1 represented the relationship between the outcome and the number of authors, β_2 represented the date of publication (mean-centered), β_3 represented if the review indicated it was funded, β_4 represented whether the review discussed multiplicity, β_5 was whether the authors adjusted the alpha rate, β_6 represented the number of studies included in the review, β_8 was the number of synthesis split reasons, β_9 was the

total number of independent syntheses, β_{10} and β_{11} were dummy variables where observational was the reference group, β_{12} and β_{13} were dummy variables where no update was the reference group, and r_i represented the error associated with each study.

The multiple regression model was an ideal way to investigate the impact of the review's characteristics on the number of statistical significance tests for a number of reasons. First, the model inherently controls for the other independent variables simultaneously. Second, conducting a multiple regression model, relative to conducting independent t-tests or ANOVA models, decreased the rate of multiplicity *in this study*. Given that the purpose of this project was to investigate and potentially decrease the number of statistical significance tests, it seemed important to follow a similar logic when conducting significance tests. Third, the simultaneous and parsimonious nature of the model engendered an ease of interpretation.

Phase II

Sample

To answer the second research question, this author also selected a small portion of the included studies from Phase I (n = 4) for further coding. This author first coded the included reviews for the presence of reported hypothesis tests' exact p-values. To be included in the sampling frame of Phase II, the review must have detailed all hypothesis tests and exact p-values or test statistics that could be turned into exact p-values. Because a limited number of studies included this information, a random selection procedure was not possible. Instead, this author purposively selected studies to illustrate the use of multiplicity corrections.

Inclusion/Exclusion Criteria

The small proportion of studies selected for further examination must also meet inclusion criteria. Specifically, the review must have detailed how the author conducted each hypothesis test and provided the exact p-values or equivalent (i.e., z-statistic, Q-statistic, etc.). Reviews that provided only statistically significant hypothesis tests were eliminated from this sampling frame. In addition, some studies reported only that the p-value was less than a certain threshold (i.e., $p < .05$). These studies were also eliminated from potential additional screening. Only studies that provided a full report of their analysis procedures received consideration.

Measurement

The subset of studies selected required more detailed coding. This stage's codebook extracted the specific hypothesis tests conducted by each review (Appendix D). Each hypothesis test included a code for the type of test (Z, Q, or other), a description of the test, the test's group, the exact test statistic, the degrees of freedom, and the p-value. No study reported all of this information for each hypothesis test; however, the p-value or the test statistic was sufficient for inclusion.

Analysis

The purpose of the subset analysis was to provide detailed information on how the reviews represented hypothesis testing and the extent to which multiplicity corrections changed the results within a given review. The first challenge was to devise a methodology for grouping the significance tests because several possibilities existed. Several scenarios were considered prior to the presentation of the final method.

Using the *a priori* grouping of significance tests, this author utilized a number of multiplicity corrections recommended by meta-analysis and primary research methodologists. This author used the Bonferroni correction, Holm's modified Bonferroni correction, Sidak's procedure (which supplanted Scheffe's procedure because it was unavailable), and the FDR procedure. The number of null hypotheses rejected following the corrections was then tabulated. This number was compared to the results of uncorrected hypothesis tests to understand how the results and conclusion may change given multiplicity corrections.

CHAPTER FOUR

RESULTS

The results have been divided into two sections. The first section detailed phase I of the project. From this dataset, this author answered research question 1. The second section detailed the analysis of the phase II. The more detailed coding and analysis derived from these studies allowed this author to answer research question 2.

Phase I

Sample

The author scanned and screened every article from every issue of PB and RER from 1986 – 2011. The process began by coding each title and abstract for inclusion. From this dataset of citations, this author proportionally randomly selected articles for inclusion, the dataset stratified by publication range (i.e., 1986-1999; 2000-2011) and publication type.

Table 2. Search and Retrieval Process

	Total Citations Screened	Met Inclusion	Randomly Selected
PB			
1986-1999	812	121 (31.6)	41 (31.6)
2000-2011	584	163 (42.6)	55 (42.6)
RER			
1986-1999	223	40 (10.4)	14 (10.4)
2000-2011	229	59 (15.4)	20 (15.4)
Total	1858	383	130

Note: Numbers in parentheses represent column proportions.

The results of this search and selection process are detailed in Table 2. A total of 1,858 citations were screened. Citations from PB constituted the largest proportion because the journal publishes up to eight issues per year while RER produces only 4 issues per year. PB published 1,396 citations over the course of 25 years and 284 of those met the inclusion criteria. RER published 452 articles from 1986-2011 and 99 were considered quantitative meta-analyses. As a proportion of the total articles published, meta-analyses represented about 20% of the total across each source (PB = 20.3%, RER = 21.9%).

To select randomly based on the proportion of the total, the column proportions must be calculated. PB citations represented 74.2% of the total and therefore were selected with the same proportionality. To ensure proportionality by publication date, the author selected the greatest proportion of studies from 2000-2011 within PB (42.6%). The second highest proportion derived from PB within the 1986-1999 year range (31.6%). Because RER produced nearly 50% less than PB, only 10.4% and 15.4% of the total citations selected derived from the ranges 1986-1999 and 2000-2011, respectively, a smaller number of RER citations were selected.

It should be noted that, on occasion, citations coded at the screening phase were coded incorrectly. Most often the study's authors seemed to indicate that a meta-analysis was conducted but upon further examination, the review utilized a narrative or vote-counting technique. In this event, the citation was coded correctly and the totals adjusted accordingly. However, this issue occurred without regularity and therefore this author considered the issue negligible. The alternative was to screen each article in its entirety;

the cost of conducting such a thorough screening process was considered to great given the benefit of eliminating a small number of false positives. As a precaution, this author randomly selected 1% (n = 14) of the rejected citations. None of randomly selected citations were coded incorrectly. This author, therefore, considered the screening process a success.

A total of 130 articles were selected for full review. This differed from the hypothesized number of 94 for one reason: Time. After the first 94 articles were coded, a significant amount of time remained relative to the project's deadline. To increase the precision of the results, this author decided to invest in coding more articles. Proportionality was maintained across the strata by randomly selecting articles that would maintain column percentages. The screening and coding process lasted roughly five months.

Descriptive Overview

Of the 130 articles selected for review, 96 (73.8%) were published in PB (see Table 3). These articles had an average of 2.65 authors and received funding 53.1% of the time. The majority of reviews in PB synthesized observational effect sizes (74.0%); only 19.8% of reviews synthesized experimental studies. Accordingly, 36.5% of reviews synthesized studies that measured differences among groups of people (i.e., male vs. female) and 27.1% estimated bivariate relationships of constructs. PB reviews utilized an efficacy purpose 10.4% of the time but rarely used an effectiveness purpose (1.0%). A large portion of studies (20.8%) attempted to predict a construct using either observational or experimental primary studies.

Thirty-four studies constituted the RER sample. On average, 3.15 authors were represented and 47.1% of reviews indicated that some type of funding was received. A large proportion of reviews synthesized experimental studies (70.6%) compared to observational studies (26.5%). A near majority of studies utilized an efficacy model of synthesis (47.1%) while only 1 study synthesized difference effect sizes (2.9%). The disparity in proportions for review purpose between the two publication sources highlighted the major differences across the two journals. PB focused more on observational studies while RER published reviews that investigated intervention effectiveness.

Table 3. General Characteristics of Included Reviews

	Total		PB		RER	
	N	%	N	%	N	%
Number of Authors	2.78	-	2.65	-	3.15	-
Funded	67	51.5	51	53.1	16	47.1
Type						
Experimental	43	33.1	19	19.8	24	70.6
Observational	80	61.5	71	74.0	9	26.5
Other	7	5.4	6	6.3	1	2.9
Purpose						
Differences	36	27.7	35	36.5	1	2.9
Bivariate	31	23.8	26	27.1	5	14.7
Efficacy	26	20.0	10	10.4	16	47.1
Effectiveness	9	6.9	1	1.0	8	23.5
Prediction	23	17.7	20	20.8	3	8.8
Other	5	3.8	4	4.2	1	2.9

Notes: Total $N = 130$; Psychological Bulletin $n = 96$ (73.8%); Review of Educational Research $n = 34$ (26.2%)

Technical aspects relating to multiplicity were also coded (Table 4). Of interest to the subset analysis, the author coded the reporting of p-values; PB reviews reported the exact p-value for all statistical tests 22.9% of the time while RER reported them only

14.1%. Somewhat surprisingly, very few of the reviews discussed the issue of multiplicity at any point in the review. In fact, only 6.9% of all reviews (n = 9) discussed the issue of Type 1 errors as a possible problem. All 9 of these reviews, furthermore, derived from PB; exactly 0 of the 34 included RER studies discussed multiplicity in any form.

Table 4. Aspects Relating to Multiplicity

	Total		PB		RER	
	N	%	N	%	N	%
Report p-values	27	21.0	22	22.9	5	14.7
Discuss multiplicity	9	6.9	9	9.4	0	0
Multiplicity Correction	25	19.2	18	18.8	7	20.6
Technique						
Implied	14	56.0	10	55.6	4	57.1
Bonferroni	7	28.0	6	33.3	1	14.3
Scheffe	2	8.0	1	5.6	1	14.3
Other	2	8.0	1	5.6	1	14.3
Power analysis	3	2.3	3	3.1	0	0
Update						
Yes	34	26.2	21	21.9	15	44.1
No	79	60.8	64	66.7	13	38.2
Partial	17	13.1	11	11.5	6	17.6

Notes: Total N = 130; Psychological Bulletin n = 96 (73.8%); Review of Educational Research n = 34 (26.2%)

While multiplicity may not have been discussed directly, review authors attempted to correct for multiple statistical tests at least a portion of the time. Of the 130 studies, 25 total studies corrected for Type 1 errors in some manner. The sources differed little with respect to the proportion of reviews that used corrections (PB = 18.8%, RER = 20.6%). Unusually, the most common correction for Type 1 error was not actually a formal correction; instead, the most commonly utilized correction was simply to increase the alpha to a nominal level below .05. The label “implied” was chosen because the

authors would rarely discuss the alpha. Instead, results tables indicated that significant results were considered below $p < .01$ instead of the traditional $p < .05$. 10 of the 18 PB reviews and 4 of the 7 RER reviews used an "implied" technique. Aside from the Bonferroni correction, used by 6 PB studies, no other correction technique was utilized more than once.

Table 5. Other Methodological Characteristics

	Total		PB		RER	
	N	%	N	%	N	%
Grey Literature	98	75.4	69	71.9	29	85.3
Study Quality	32	24.6	16	16.7	16	47.1
Cohen's classification	64	49.2	49	51.0	15	44.1
Model specification						
No	26	20.0	10	10.4	16	47.1
Fixed and Random	10	7.7	8	8.3	2	5.9
Fixed	15	11.5	10	10.4	5	14.7
Random	39	30.0	28	29.2	11	32.4
Graphical plot	62	47.7	44	45.8	18	52.9
Publication bias*						
None	55	42.3	40	41.7	15	44.1
Fail-safe N	29	22.3	23	24.0	6	17.6
Moderator test	20	15.4	14	14.6	6	17.6
General funnel plot	7	5.4	3	3.1	4	11.8
Egger's test	7	5.4	7	7.3	0	0
Trim and Fill	12	9.2	10	10.4	2	5.9
Other	2	1.5	1	1.0	1	2.9

Notes: Total $N = 130$; Psychological Bulletin $n = 96$ (73.8%); Review of Educational Research $n = 34$ (26.2%); Publication bias total does not add to 100% because some studies used multiple measures.

One important code to the multiplicity literature in meta-analysis was whether the review constituted an update. A review was considered a partial update if the review authors indicated that some portion of the review overlapped with an existing review. 11.5% and 17.6% of the PB and RER reviews were considered partial, respectively.

Interestingly, only 38.2% of RER reviews were unique reviews (i.e., no overlap with any existing review). PB published reviews were most often unique, on the other hand (66.7%).

Finally, this author coded a number of methodologically interesting aspects. Inclusion of grey literature constituted at least some recognition from review authors (Table 5). A large majority of reviews included some form of grey literature (PB = 71.9%, RER = 85.3%). This author did not code the type of grey literature. More RER review authors assessed primary study quality (47.1%) relative to PB reviews (16.7%). RER often failed to report model specification, however, relative to PB reviews (PB = 10.4%, RER = 47.1%). One major area of interest as of late, furthermore, has been on publication bias. To investigate this paradigm, this author coded what type of method the review authors utilized to measure potential publication bias. The results indicated that a majority of PB reviews utilized the "Fail-safe N" (24.0%) or simple moderator test of publication status (14.6%). RER reviews utilized the "Fail-safe N" and moderator tests less frequently, using them 17.6% each, respectively. Surprisingly, most studies failed to report any type of publication bias indicator. 41.7% of PB reviews and 44.1% of RER reviews did not report publication bias indices.

Statistical Tests Usage

To answer the primary research question, this author coded and analyzed the number and type of statistical test usage across the sample (Table 6). To clarify, each test of statistical significance was coded, including the overall test of statistical significance, the overall test of heterogeneity, tests of subgroup average effect size, each moderators

test of homogeneity Q-test (including Q-within and Q-between tests), pairwise comparison tests, and all meta-regression tests of slopes. This author did not count statistical significance tests labeled as "sensitivity analyses" or correlations among the study characteristics. In addition, hypothesis tests that did not report or use significance tests or used confidence intervals only were not counted as having used tests of statistical significance. The results revealed that, across both journal and all years, review authors utilized an average of 70.82 tests of statistical significance (*s* = 94.20; Median [M] = 46.5).

Table 6. Statistical Test Usage

	Total			PB		RER	
	Mean	Median	Range	Mean	Median	Mean	Median
Statistical tests	70.82 (94.20)	46.5	0-920	77.53 (105.51)	50.5	51.85 (49.17)	36.5
Tests of overall ES	7.15 (19.48)	1.00	0-178	7.17 (14.00)	1.50	7.09 (30.32)	1.00
Tests of Homogeneity	9.14 (14.94)	3.00	0-72	11.43 (16.71)	4.00	2.69 (3.34)	1.50
Q-between	9.61 (17.79)	0.00	0-105	9.11 (18.38)	0.00	11.00 (16.16)	0.00
Q-within	13.51 (29.94)	0.00	0-194	16.79 (33.68)	0.00	4.24 (10.94)	0.00
Z-tests (as moderators)	20.51 (84.12)	0.00	0-920	21.64 (95.75)	0.00	17.32 (35.49)	0.00
Meta-regression tests	10.90 (27.53)	0.00	0-235	11.39 (29.71)	0.00	9.53 (20.50)	0.00
Syntheses	12.72 (21.26)	5.00	1-178	13.86 (17.19)	6.00	9.47 (30.01)	3.00
% More than 1 synthesis	80.77	-	-	81.25	-	79.41	-

Notes: *N*s: Psychological Bulletin = 96, Review of Educational Research = 34; Numbers in parentheses represent SD.

One prominent reason for the high rate of statistical significance was the number of "independent" syntheses per study. The average review conducted 12.72 independent

syntheses (s = 21.26, M = 5.00). Moreover, 80.77% of reviews conducted more than one synthesis. The reason for these splits will be explored further in the next section of this chapter. Of those syntheses, only little more than half conducted a corresponding test of overall significance (μ = 7.15, s = 19.48, M = 1.00). Many studies simply reported the 95% confidence interval and these were not counted as tests of statistical significance. Each independent synthesis more often conducted a corresponding tests of overall homogeneity (μ = 9.14, s = 19.48, M = 3.00).

Overall tests of significance represent only one type of statistical test. Reviews, on average, also conducted 9.61 moderator tests of heterogeneity (i.e., Q – Between; s = 17.79, M = 0.00). In addition, reviews routinely conducted multiple tests of Q – Within (μ = 13.51, s = 29.94, M = 0.00). A large number of z-tests used for the purpose of moderator analyses were also conducted (μ = 20.51, s = 84.12, M = 0.00). As a point of clarity, tests of significance labeled as moderator z-tests were those conducted for the purpose of either subgroup analyses or multiple comparisons. For example, Hostetter (2011) tested the overall effect size significance, synthesizing 63 independent studies. To analyze these studies further, the author categorized the effect sizes into various groups, such as age or population. The author then tested the overall significance of each groups' effect size. In contrast, Connelly and Ones (2010) conducted multiple comparison tests by grouping effect sizes into independent groups (i.e., each effect size could only belong to one category) before testing whether the difference between each group was significant. Although this constitutes two unique philosophical approaches to moderator

analyses, the statistical test remained the same. Therefore, each significance tests was labeled as a moderator z-test.

Table 7. Statistical Test Usage by Source and Year

	1986-1999		2000-2011	
	Mean	Median	Mean	Median
N statistical tests				
PB	51.71 (51.84)	34.0	96.78 (128.61)	71.0
RER	59.93 (61.08)	32.0	46.20 (39.55)	39.0
N Tests of overall ES				
PB	5.24 (11.46)		8.61 (15.58)	
RER	15.21 (46.95)		1.40 (2.19)	
N Tests of Homogeneity				
PB	7.14 (12.10)	2.00	14.62 (18.92)	6.00
RER	2.43 (3.16)	1.50	2.85 (3.53)	1.50
N Q-between				
PB	6.76 (12.91)	0.00	10.87 (21.53)	0.00
RER	10.86 (13.49)	6.50	11.10 (18.14)	3.50
N Q-within				
PB	16.14 (39.86)	0.00	17.27 (28.59)	0.00
RER	2.00 (7.48)	0.00	5.80 (12.77)	0.00
N Z-tests (as moderators)				
PB	8.44 (19.17)	0.00	31.47 (124.99)	0.00
RER	20.42 (34.18)	0.00	15.15 (37.10)	0.00
N meta-regression tests				
PB	7.98 (16.01)	0.00	13.93 (36.72)	0.00
RER	9.00 (22.21)	0.00	9.90 (19.79)	0.00
N syntheses				
PB	10.54 (13.46)	5.00	16.35 (19.26)	7.00
RER	16.00 (46.71)	3.50	4.90 (4.23)	3.00
% More than 1 synthesis				
PB	78.05	-	83.64	-
RER	78.57	-	80.00	-

Notes: Ns: PB 1986-1999 = 41, PB 2000-2011 = 55, RER 1986-1999 = 14, RER 2000-2011 = 20; Numbers in parentheses represent SD.

Finally, this author also coded and counted the number of meta-regression tests of significance. Again, a point of clarity should be provided. Only tests of significance that were conducted between a moderator (i.e., percentage of female participants) and an

effect size. This eliminated tests of significance that tested the correlation among the moderators. The results indicated that 10.90 meta-regression tests of significance were conducted per study (s = 27.53, M = 0.00).

Given their disparate nature and sampling process utilized, this author decomposed the number of statistical tests by the publication sources (Table 7). Reviews from PB conducted 33.1% more tests of statistical significance relative to RER reviews (μ: PB =77.53, RER = 51.85). A breakdown of the type of test revealed large differences in the number of Q-within tests of statistical significance (μ: PB = 16.79, RER = 4.24). It is difficult to hypothesize the reason for this large disparity; PB reviews might simple decompose the "levels" of studies with greater propensity. Another reason could be that RER studies failed to report this statistical test. PB reviews also tended to include a greater number of independent syntheses (μ: PB = 13.86, RER = 9.47).

To dissect further the dimensions of the sampling frame, this author decomposed the number of statistical tests by the source and year as well (Table 7). The outcome of primary interest, total number of statistical tests per study, changed drastically for PB reviews. During the time period of 1986-1999, PB reviews conducted 51.71 tests of statistical significance (s = 51.84, M = 34.0). Over the eleven year period of 2000-2011, however, reviews published within PB conducted almost double the amount of statistical tests (μ = 96.78, s = 128.61, M = 71.0). Over that same time period, the average number of statistical tests conducted within RER reviews *decreased* (μ: 1986-1999 = 59.93, 2000-2011 = 46.20). This decrease was well within the expected range and could easily be attributed to statistical variation given the smaller sample size (n = 34). Moreover, the

median number of statistical tests increased (M: 1986-1999 = 32.0, 2000-2011 = 39.0). Most likely the number of statistical tests remained unchanged over the course of those years. The same should not be hypothesized for PB reviews; the number of statistical tests almost certainly increased over the past decade relative to the two previous decades.

The increase in statistical tests in PB reviews can be attributed to the increased number of z-tests for moderators. From 1986-1999, PB reviews averaged 8.44 z-tests for moderators (s = 19.17). From 2000-2011, PB reviews increased the average number of z-tests for moderators to 31.47 (s = 124.99). That represented an increase of 380% more statistical tests of significance over the last decade. RER reviews, in contrast, again decreased the number of statistical significance tests, averaging 20.42 from 1986-1999 and 15.15 from 2000-2011. The other types of tests remained relatively unchanged and within the expected statistical range.

The number of independent syntheses also revealed interesting results. Reviews published in PB increased the number of independent syntheses from 10.54 to 16.35. This represented a modest increase in the number of syntheses. In contrast, reviews published in RER diminished the number of independent syntheses drastically, decreasing the number of syntheses from 16.00 to 4.90. This represented a near 70% drop in the total number of syntheses. Again, these changes should be considered expected as the median number of syntheses decreased only slightly for each group.

Relationships among Statistical Tests. One way to investigate further the nature of statistical significance testing is to conduct correlational analyses of the relationships

between the types of tests. This author conducted simple bivariate correlations at the study-level for each of the 6 significance testing variables (Table 8).

Table 8. Relationships among the Statistical Tests

	1.	2.	3.	4.	5.
1. Number of Overall Z	-				
2. Number of Overall Q	.31**	-			
3. Number of Q-Between	.01	.02	-		
4. Number of Q-Within	-.14	-.13	.35**	-	
5. Number of Z-Tests of Moderator	-.05	-.05	-.07	-.05	-
6. Number of Meta-Regression	-.01	.24**	.05	-.09	-.09

Notes: $N = 130$, * $p < .05$, ** $p < .01$

The results of the correlation analysis indicated several findings of interest. The relationship between the number of overall Z tests and number of overall Q tests was positive and significant ($r = .31$, $p < .01$). The relationship between the number of Q-Between and Q-Within tests of significance was positive and significant ($r = .35$, $p < .01$), an expected finding considering that when one conducts Q-Between tests of significance Q-Within tests almost surely follow. A curious result was found for the relationship between overall Q tests of homogeneity and the number of meta-regression tests of significance ($r = .24$, $p < .01$); there was no reason to expect this relationship and should be treated with caution. Somewhat surprisingly, the relationship between the number of z-tests for moderator variables did not significant correlate with any of the other variables. This author hypothesizes that this was due to the fact that meta-analysts used these tests differently, some reviews as subgroup tests of overall significance and some in a pairwise comparison context.

Multiple Syntheses in One Review

A meta-analyst must decide to "split" or "lump" effect sizes from studies when conducting the final analyses. Bender et al. (2008) hypothesized several reasons why an

author would consider splitting effect sizes. To reiterate, Bender et al. hypothesized that a meta-analysis could split effect sizes due to different types of outcomes, type of effect sizes, groups from the populations (i.e., treatment, comparison, etc.), or time points.

While these four reasons for a meta-analysis split indeed constituted reasons to split a meta-analysis, this author quickly realized that they were insufficient to represent the different reasons meta-analysis authors chose to split up effect sizes. This section will detail the newly recognized reasons to split up effect sizes due to heterogeneity, provide an analysis of the reasons for splitting the sample, and detail two examples of what occurs when a meta-analyst determines that effect sizes must be split.

Description of Results. The reasons provided by Bender et al. (2008) were found to be insufficient to represent the myriad reasons for splitting effect sizes into independent syntheses. Table 9 delineated both the old conceptualization (i.e., Bender et al.'s) as well as this author's findings. Before describing these categories, it is important to note that these categories do not constitute the totality of possibilities. Given that the sample consisted of education and psychology studies, these categories of heterogeneity may simply represent the common reasons to split a sample across this specific literature. Nevertheless, this categorization represented a unique development.

The new reasons for meta-analysis splits have been presented in the second half of the table. The table has four columns: Source, Reason for split, Description, and Example from sample. This author coded five new reasons for splitting a meta-analysis (not including an "other" category). The first new reason represented in the table constituted a split by the predictor or subscale. For example, Eagly and Johnson (1990)

synthesized the gender difference in leadership styles. The authors grouped the effect sizes according to predetermined subscales that represented divergent constructs of interest. In total, Eagly and Johnson's meta-analysis synthesized five independent sets of effect sizes. The authors stopped conducting analyses after the synthesis thereby not conducting moderator analyses.

A second newly defined reason represented the comparison group of interest. Although Bender et al.'s (2008) representation of synthesis splits introduced treatment group as a potential reason, this limited the splits purely to experimental studies. Observational studies, as well as differing control-group studies, required a different split representation. The comparison group split occurred when a study used multiple populations as the comparison of interest. A unique example of this type of split was presented by Grabe and Hyde (2006). This review synthesized women's body dissatisfaction as a function of their ethnicity. To estimate the average effect sizes, the authors calculated a mean-difference on a body dissatisfaction for multiple ethnicity comparisons: White-Black, White-Hispanic, White-Asian American, Black-Hispanic, Black-Asian American, and Hispanic-Asian American. The authors then conducted moderator analyses for each group.

A third example of a new type of defined split derived from the populations synthesized. This follows a similar logic of the previous example, however, the primary studies represented within the independent syntheses are generally assumed to have a homogenous population. Sporer, Penrod, Read, and Culter (1995) synthesized studies that conducted lab-controlled experiments that divided participants into choosers or non-

choosers. The authors operationalized choosers as individuals who correctly identified persons of interest while non-choosers chose not to identify a person of interest. The percentage of choosers and non-choosers was calculated for each primary study; a study with a majority of choosers would be labeled as a chooser study. Thus two separate, independent, syntheses were conducted within this single review: Primary studies with choosers and studies with non-choosers. Separate moderator analyses were also conducted for each of the independent syntheses.

The source of the data for the primary study was also rationalized as reason to split syntheses. The rational generally presented was that National datasets represented too large a sample to fairly be included with researcher-selected samples. Hyde, Fennema, and Lamon (1990) synthesized gender differences in mathematics performance across a variety of mathematics outcomes. The review authors split the effect sizes by SAT and non-SAT samples, rationalizing that primary studies using SAT scores "exerted a disproportionate effect" (pg. 146). An overall average effect was calculated for each independent syntheses yet moderator analyses were conducted only for the non-SAT sample.

A final reason for synthesis split was given via a statistical reasoning. This represented a broad category of reasons as opposed to the specific categories previously presented. Here, the meta-analyst determined that effect sizes should be represented differently given an unique statistical issue. The most common statistical reasoning provided was the "level of effect". Primary studies may report an effect size at any number of hierarchical levels (Raudenbush & Bryk, 2002). For example, Swanson and

Hoskyn (1998) synthesized effect sizes both at the study-level and the individual effect size level.

Table 9. Old and Updated Reasons for Synthesis Splits

Source	Reason for split	Description	Example
Bender et al. (2008)	Outcome	The outcomes of interest are deemed to divergent to synthesize.	Oliver & Hyde (1993)
	Effect size	The effect sizes chosen to represent each study are not amenable to synthesis (i.e., pre-post d and treatment-control d).	Kling, Hyde, Showers, & Buswell (1999)
	Treatment group	An efficacy study synthesizes multiple treatments but represent different philosophical constructs.	Bowers, Kirby, & Deacon (2010)
	Time point	A review synthesizes effect sizes from post-test and follow-up observations.	Ross (1988)
New	Predictor/Subscale	The review synthesizes correlations between one outcome and multiple types of other variables; the review synthesizes multiple subscales of one construct.	Eagly & Johnson (1990)
	Comparison group	A review determines that the control groups (or contrasting groups) represent different types of people.	Grabe & Hyde (2006)
	Participant type	Samples included in the primary study represent different populations.	Sporer, Penrod, Read, & Cutler (1995)
	Data source	Synthesizes are grouped by the data source (i.e., National samples are excluded from primary analyses).	Hyde, Fennema, & Lamon (1990)
	Statistical reason	Synthesizes are split for some statistical reasons given by the author (i.e., level of primary study analysis, synthesis model).	Swanson & Hoskyn (1998)

A final group, “other”, was not represented in the table. The reasons indicated in the other group were type of survey, type of measure, and type of validity indicator (i.e.,

construct, content, etc.). These reasons for synthesis splitting failed to fall broadly within one of the other nine categories and were therefore not included as reasons.

Example of Multiple Splits. The studies presented as examples represented reviews where a small number of splits occurred. To better understand the way reviewers split syntheses, a visual representation of the splits has been created for two examples. The first example, McCartney, Harris, and Bernieri (1990), synthesized twin studies that compared trait differences within each dyad. Figure 3 indicated how the effect sizes were split by the meta-analysts to form independent syntheses. This review used two "splits" to form the independent syntheses. First, the authors split the effect sizes into the traits of interest; for this study, the authors utilized 11 traits (outcomes). (Note that not all outcomes were represented in the figure.) Next, the authors further divided the effect sizes by the participant type, in this case, the type of twin dyad. Each outcome had only two types of twins, so each outcome provided two independent effect sizes, one for each type of twin study. This review, therefore, conducted 22 independent syntheses.

The authors then conducted three types of statistical tests. The first tested the overall correlation's slope (i.e., H_0: $r = 0$). The second tested for homogeneity of the effect sizes within each independent study (i.e., H_0: $Q = 0$). The third was a test of three moderators, age, type of report, and zygosity determination. Again, these moderator analyses were conducted for each independent synthesis; however, each statistical analysis was not reported. In sum, the authors conducted 110 tests of statistical significance.

Miller, Turner, Tindale, Posavac, and Dugoni's (1991) also split the sample of effect sizes into independent syntheses, but used a far greater number of reasons (Figure 4). In contrast to the previous example, only one outcome was synthesized. The first split occurred because of the type of data source, either prospective or cross-sectional. A second split divided each of the two groups into the types of populations, high risk or healthy. A third split partitioned the cross-sectional, healthy population into two more groups: Case controlled studies or population-based. Finally, a fourth split occurred for the type of survey that the primary studies utilized. A total of 14 independent syntheses were conducted. The review authors choose not to conduct moderator analyses which limited the number of statistical tests within the study. Only two types of statistical tests were utilized, the test of the correlation and a test of the synthesis heterogeneity. This review, therefore, conducted only 28 null hypothesis significance tests.

Figure 3. McCartney et al. (1990) Splitting Example

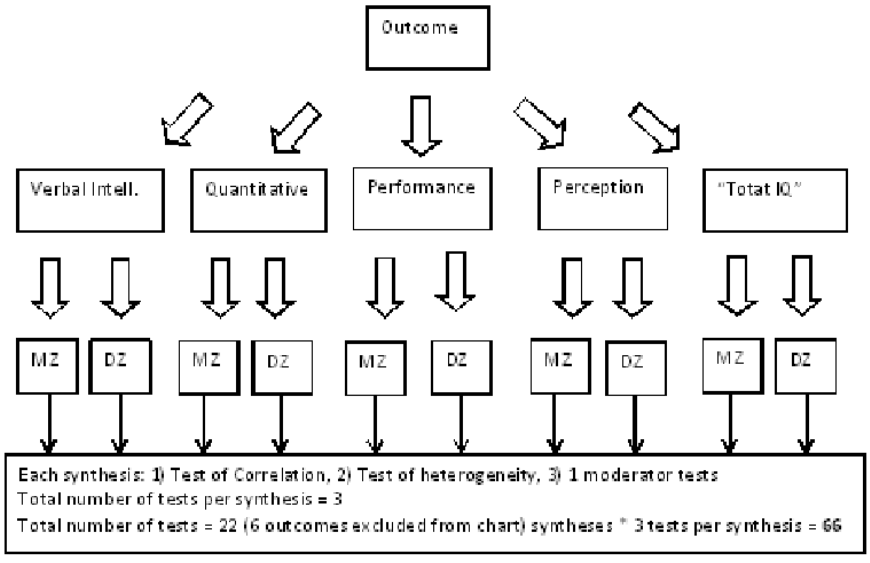
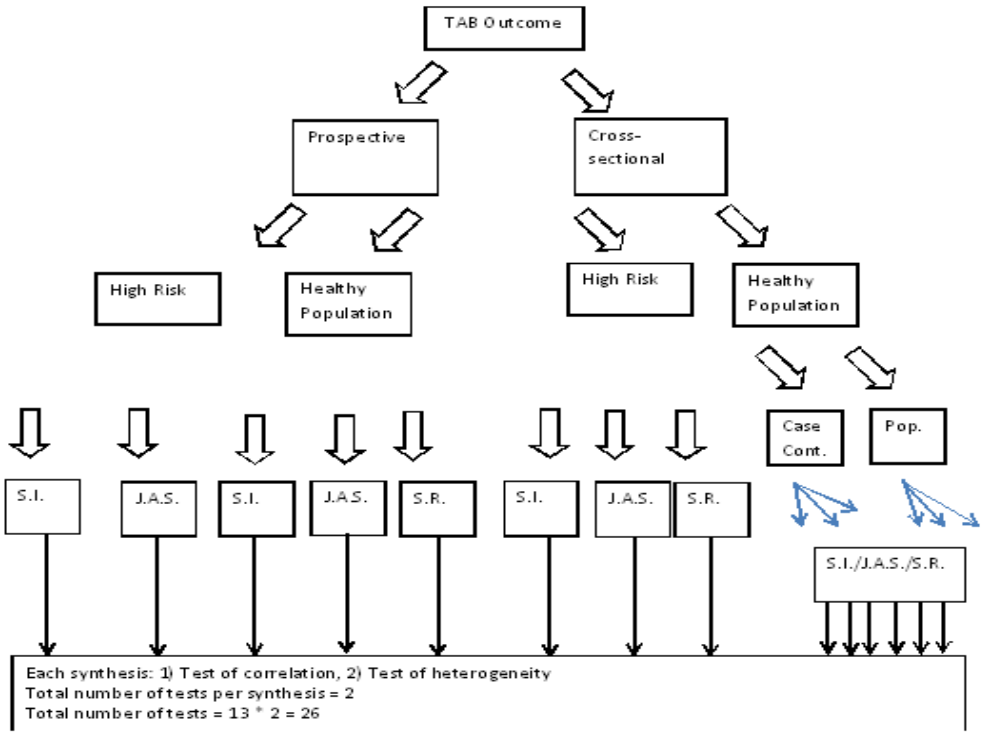


Figure 4. Miller et al. (1991) Splitting Example



Occurrence and Reasons for Splitting. The occurrence of synthesis splitting remains unanswered across the psychology and education disciplines. To answer this specific question, the number of splits was recorded. The results indicated that 25 of the 130 (19.23%) syntheses "lumped" all the effect sizes into a single average effect. More than 4 out of every 5 review, however, subdivides the effect sizes into independent syntheses prior to calculating an average effect size.

Another important question to answer, then, is the rational given for synthesis splitting. Table 10 indicated the row percentage for each of the 9 reasons (plus a 10th reason for "other"). Forty-nine of the one-hundred and thirty studies included in the

review split the effect sizes only once. Of those 49, 21 (42.86) studies indicated that the synthesis split was due to outcome differences. The second highest reason indicated for studies with one split was because of predictor or subscale divergence (16.33%). In fact, for reviews that split 1-3 times, the top two reasons for splitting were always outcome and predictor/subscale, respectively. Reviews with more than 4 splits behaved slightly differently, but only 6 studies represented the entirety of these types.

The findings were also analyzed at study-level for further investigation (Table 11). Of the 105 studies that had at least one split, 68 reviews listed one of the reasons as outcome related (64.76%). The second most used reason for splitting the sample was due to the predictor or subscale (31.43%). The third most reason was the participant type (16.19%). Across all studies included in the review, 52.3% split the effect sizes by outcome.

Table 10. Reasons for Splitting the Effect Size by Number of Splits

	N	O	ES	P/S	TG	CG	PT	DS	TP	SR	OT
1 Split	49	42.86%	0.00%	16.33%	6.12%	10.20%	4.08%	2.04%	2.04%	2.04%	0.00%
2 Splits	38	43.42%	3.95%	21.05%	3.95%	3.95%	7.89%	3.95%	2.63%	6.58%	2.63%
3 Splits	13	25.64%	5.13%	15.38%	2.56%	5.13%	17.95%	12.82%	2.56%	7.69%	5.13%
4 Splits	3	25.00%	25.00%	16.67%	0.00%	0.00%	0.00%	8.33%	0.00%	25.00%	0.00%
5 Splits	1	0.00%	20.00%	0.00%	0.00%	20.00%	20.00%	20.00%	20.00%	0.00%	0.00%
6 Splits	0	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
7 Splits	1	14.29%	14.29%	14.29%	0.00%	14.29%	14.29%	0.00%	14.29%	14.29%	0.00%

Notes: N = Sample size, O = Outcomes, ES = Effect size, P/S = Predictor/Subscale, TG = Treatment group, CG = Comparison group, PT = Participant type, DS = Data source, TP = Time point, SR = Statistical reason, OT = Other.

Table 11. Reason for Splits

Reason for Split	Number of Reasons	Percentage of Total Studies with Splits (N = 105)	Percentage of Total Studies (N = 130)	Percentage of Total Reasons (N = 181)
Outcome*	68	64.76%	52.31%	37.57%
Effect Size*	10	9.52%	7.69%	5.52%
Predictor/Subscale	33	31.43%	25.38%	18.23%
Treatment Group*	7	6.67%	5.38%	3.87%
Comparison Group	12	11.43%	9.23%	6.63%
Participant Type	17	16.19%	13.08%	9.39%
Data Source	11	10.48%	8.46%	6.08%
Time Point*	6	5.71%	4.62%	3.31%
Statistical Reason	13	12.38%	10.00%	7.18%
Other	4	3.81%	3.08%	2.21%

Notes: Percentages do not add to 100% because studies indicated more than one split reason; * indicate reasons given by Bender et al. (2008).

Interestingly, the three examples provide by Bender et al. (2008) were rarely cited as reasons for splitting syntheses. For syntheses with only one split, the treatment group was provided as the reason only 6.12% of the time, time point 2.04%, and effect size metric was never given as a reason. Reviews that split syntheses twice also used the effect size, treatment group, and time point infrequently (3.95%, 3.95%, and 2.63%, respectively). Meta-analyses that had three reasons to split the effect sizes, a total of 13 studies, also split the effect sizes at a similar rate for those three reasons (5.13%, 2.56%, and 2.56%, respectively). At the study-level, effect size, treatment group, and time point ranked as the 7th, 8th, and 9th given reasons, respectively. Clearly the claims made by Bender et al. were unwarranted or, at the least, misplaced.

Of course, the reason for the lack of splitting could be due to several factors. The most likely reason remained that the authors eliminated effect sizes, or simply did not include effect sizes, that represented a difference in effect size, treatment group, or time point. Another reason, more relevant to the time point reason, was that primary studies only collected post-test observations and did not collect follow-up observations. Whatever the reason, Bender et al.'s (2008) rational for splitting syntheses failed to materialize for reasons other than outcome splitting.

Synthesis Splitting and the Number of Independent Syntheses. Inherently, as the number of reasons for splitting the effect sizes increases, the number of independent syntheses increases. To investigate and estimate this relationship quantitatively, however, this author estimated the number of independent syntheses per the number of reasons for splitting the effect sizes (Table 12). In concordance with expectation, the number of

syntheses increases linearly as a function of the number of reasons for splitting. The numbers listed in the table, as a point of clarity, were the number of independent syntheses included within one review and not the statistical significance tests.

Table 12. Number of Independent Syntheses by the Number of Split Reasons

Number of Reasons	Overall	PB	RER
0 Reasons (n = 25)	1.00 (0.0)	1.00 (0.0)	1.00 (0.0)
1 Reasons (n = 49)	6.49 (8.21)	8.03 (9.75)	3.59 (2.06)
2 Reasons (n = 38)	16.45 (16.20)	18.48 (17.22)	7.43 (4.53)
3 Reasons (n = 13)	26.15 (20.81)	28.73 (21.73)	12.00 (0.0)

Notes: Numbers in parentheses represent SD.

Reviews where only 1 reason was provided for splitting synthesized, on average, synthesized 6.49 different groups of effect sizes. Reviews where two reasons were given for splitting synthesized 16.45 average effect sizes. This number continued to increase for reviews that included 3 reasons for splitting effect sizes ($\mu = 26.15$, $s = 20.81$). The results also revealed that this varied as a function of the source. Not surprisingly in light of the previous analyses, PB included a greater number of independent syntheses relative to RER. It should also be noted that studies with more than 3 synthesis splits ($n = 5$) were not included in this analysis. Simply stated, there were not enough studies represented in those categories to estimate precisely the average number of syntheses. One would expect, however, that syntheses with a large number of splits (i.e., splits > 3) will in turn utilize a greater number of syntheses. This hypothesis simply needs greater investigation.

Splitting and Statistical Tests. To observe how the number of splits affected the number of statistical tests, this author also calculated the number of statistical tests and type of tests by the number of splits (Table 13). Results again were confined to reviews where the number of reasons was equal to or less than 3 splits.

The results confirmed the inherent problem in splitting the effect sizes into independent syntheses. The syntheses that lumped all the effect sizes into a single analysis (*n* = 25) conducted, on average, only 41.32 tests of statistical significance (*s* = 39.99). In fact, the number of statistical tests increased at each level of splitting. Studies that split syntheses only once (*n* = 45) used 62.32 tests of statistical significance; syntheses with exactly two splits (*n* = 38) conducted 88.55 tests of significance. Although disproportionately under-utilized, studies that had three splits used 95.31 (*s* = 58.04) tests of statistical significance. Clearly the trend indicated that more synthesis splits will impact the number of statistical tests.

Table 13. Number of Statistical Tests by the Number of Split Reasons

	0 Splits	1 Splits	2 Splits	3 Splits
Total Number of Tests	41.32 (39.99)	62.32 (56.83)	88.55 (151.52)	95. 31 (53.94)
Total Tests of Overall Average Effect	.76 (.43)	3.25 (5.64)	5.79 (9.31)	15.54 (22.84)
Total Tests of Overall Homogeneity	.84 (.37)	4.92 (8.39)	12.76 (15.81)	24.92 (22.08)
Total Number of Q-Between Tests	6.24 (9.02)	14.71 (21.87)	7.29 (17.84)	7.31 (12.35)
Total Number of Q-Within Tests	12.48 (27.77)	19.43 (38. 57)	8.42 (15.39)	13.23 (33.31)
Total Number of Z-Tests for Moderators	14.28 (18.66)	12.88 (31.78)	35.0 (149.61)	21.85 (34.83)
Total Number of Meta-Regression Tests	6.72 (10.45)	7.15 (17.85)	19.29 (44.05)	12.46 (17.74)

Notes: *N*s = 25, 49, 38, 13, respectively. Numbers in parentheses represent SD.

Predicting the Number of Statistical Tests

The previous sections detailed descriptively, without the use of statistical significance testing, the relationships between contextual variables and significance testing. To decrease the use of statistical testing, thereby decreasing the probability of this author also committing a Type 1 error, a multiple regression model was utilized to test for

moderators of statistical test usage simultaneously. The traditional assumptions of multiple regression (i.e., outcome normality, multicollinearity, independence) were first evaluated for tenability. The distribution of the total number of statistical tests was highly positively skewed and, as a result, was transformed using a log-linear transformation. The results of the multicollinearity test revealed low to moderator correlations among the predictors (Table 14). Only one correlation remained a concern; the relationship between the number of reasons for splitting the effect sizes and the number of independent syntheses ($r = .68$, $p < .01$). Multicollinearity diagnostics, however, indicated that the covariance between these variables did not impact the model's results (VIF = 1.96). Finally, independence was handled by aggregating all information about the statistical tests to the study level; each study provided only one set of indicators and outcome information.

The independent variables for the multiple regression were chosen to represent the contextual aspects of the sample of meta-analyses (EQ. 17). The independent variables included in the model were the number of authors listed, whether the study received outside funding, if the study authors discussed multiplicity, whether the meta-analysts adjusted the alphas, and how many studies were included in the review. In addition, the model included variables related to the type of meta-analyses: the number of reasons for spits, the number of independent syntheses, the type of meta-analysis (i.e., observational, experimental, or both), and if the review was a full or partial update. The date of publication was mean-centered for ease of interpretation.

Table 14. Correlation Matrix of Variables included in the Model

	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.
1. Source	-													
2. Authors	.16	-												
3. DoP	.05	.29**	-											
4. Funded	-.05	.17*	.03	-										
5. Discuss Multiplicity	-.16	-.02	.01	.02	-									
6. Adjust Alphas	.02	-.12	-.18*	.01	.33	-								
7. N of Studies	.02	.01	.02	.08	-.10	.21*	-							
8. Reasons for Splits	-.08	.01	.01	.16	.03	.03	-.22**	-						
9. Ind. Syntheses	-.09	-.10	-.01	.06	.10	.01	-.17*	.68**	-					
10. Type = Obs.	-.43**	-.07	-.09	-.01	-.03	-.14	-.09	.16	.13	-				
11. Type = Exp.	.47**	.12	.09	-.04	-.13	.03	.11	-.28**	-.25**	-.89**	-			
12. Type = Both	-.06	-.11	.01	.10	.34**	.23**	-.05	.24	.24**	-.30**	-.17	-		
13. No Update	-.20*	-.25**	-.32**	.01	.10	.07	-.13	-.03	.04	-.05	-.04	.19	-	
14. Full Update	.16	.28**	.17	-.02	-.02	-.07	.21*	.03	-.06	.04	.03	-.14	-.74**	-
15. Partial Update	.08	-.01	.24**	.01	.01	-.02	-.09	.01	.02	.03	.02	-.09	-.48**	-.23**

Notes: N = 130; Source: 1 = PB, 0 = RER; DoP = Date of Publication* $p < .05$, ** $p < .01$.

The results of the regression model revealed several findings of interest (Table 15). As seen previously, the date of publication was significantly related to the number of tests conducted (β .03, SE = .01, B = .21, p = .01). A study conducted in 2005, 5 years after the average year in the dataset (i.e., 2000) is predicted to conduct 5.87 more tests of statistical significance, controlling for all other variables. Also, the number of independent syntheses predicted the number of statistical tests (β = .02, SE = .01, B = .31, p = .01). In fact, the number of syntheses was the strongest predictor of the number of statistical tests.

Table 15. Predictors of Total Number of Statistical Tests

Variable	Variable Coding	β (SE)	B	95% CI	p-value
Constant	-	3.59 (.24)	-	3.04, 4.00	.01
Source	1 = RER, 0 = PB	.01 (.21)	.01	-.42, .43	.99
Authors	-	-.04 (.06)	-.06	-.16, .08	.52
Date of Publication	Mean-Centered	.03 (.01)	.24	.01, .05	**.01**
Funded	1 = Yes, 0 = No	.27 (.17)	.13	-.06, .61	.11
Multiplicity	1 = Yes, 0 = No	.48 (.36)	.12	-.22, 1.19	.18
Adjust Alphas	1 = Yes, 0 = No	.01 (.23)	.01	-.44, .47	.96
Studies	Review Total	.002 (.001)	.22	.001, .003	**.01**
Split Reasons	Review Total	.09 (.10)	.10	-.11, .29	.38
Syntheses	Review Total	.02 (.01)	.31	.01, .03	**.01**
Type = Experimental	1 = Yes, 0 = No	-.37 (.20)	-.17	-.77, .03	.07
Type = Both	1 = Yes, 0 = No	-.45 (.40)	-.10	-1.24, .34	.26
Full Update	1 = Yes, 0 = No	-.27 (.21)	-.12	-.68, .14	.20
Partial Update	1 = Yes, 0 = No	-.01 (.26)	-.01	-.53, .50	.96
R^2	-	.31	-	.19, .43	-
F	-	3.96	-	-	.001

Notes: N = 130; Total number of tests log-transformed; Type, Observational = Reference group; No Update = Reference group.

Two other findings of interest, not specifically mentioned previously, should be noted. First, the number of studies included in the review was a significant predictor of the number of statistical tests (β = .002, SE = .001, B = .22, p =.01). The results of this finding remained somewhat unsurprising given that many meta-analysis textbooks

suggest that a certain number of studies must be included in the review (i.e., 10 or more) in order to conduct statistical analyses (Lipsey & Wilson, 2001). Clearly authors believed that more studies also constituted the right to conduct more statistical tests. More surprisingly, however, observational syntheses, relative to experimental syntheses, conducted nearly significantly more tests of significance (β = -.37, SE = .20, B = -.17, *p* =.07). This finding undoubtedly arose due to the high number of bivariate correlations often included within correlational reviews. Each correlation, on average, has the potential to provide two tests of significance as well as multiple moderator analyses. Experimental meta-analyses, on the other hand, generally synthesize specific relationships or outcomes.

The model-fit statistics indicated that this was a well-fitting model ($F = 3.96$, $p < .001$). Overall, the model explained 31% of the dependent variable variance (95% CI: .19, .43). Given the high degree of variability among the types of studies, reviewer preferences, synthesis techniques, and publication date, this model provided sufficient information.

To ensure that the results of the regression model were not spurious due to multicollinearity or small sample size, this author estimated a reduced form model removing all non-significant variables (Table 16). The results of the reduced form model matched those of the full model. The strongest predictor of statistical test usage was again the number of syntheses (β = .02, SE = .01, B = -.43, *p* =.01). The date of publication (β = .03, SE = .01, B = .19, *p* =.02) and the number of studies included in the review (β = .001, SE = .001, B = .18, *p* =.03) remained significant predictors as well. Relative to the

full model, the reduced form model explained less variation in the outcome (R^2 = .22, 95% CI: .10, .34). The results of this model indicated, however, that the variables initially indicated as significant predictors remained significant predictors.

Table 16. Reduced Model Predicting the Total Number of Statistical Tests

Variable	Variable Coding	β (SE)	B	95% CI	p-value
Constant	-	3.43 (.11)	-	3.22, 3.65	.01
Date of Publication	Mean-Centered	.03 (.01)	.19	.01, .05	**.02**
Syntheses	Review Total	.02 (.01)	.43	.01, .03	**.01**
Studies	Review Total	.001 (.001)	.18	.001, .003	**.03**
R^2	-	.22	-	.10, .34	-
F	-	12.02	-	-	.001

Notes: *N* = 130; Total number of tests log-transformed.

Phase II

The results of the previous section revealed an enterprise reliant on null hypothesis testing. Such a large number of statistical significance testing inherently leads to biased claims and mitigated answers. Left unchecked, as mentioned previously, multiple false rejections of the null hypothesis within a single study is likely to occur (see Figure 2).

One solution to the problem derives from primary study statistical literature: Multiplicity corrections. There remain two inherent issues that relate to multiplicity corrections that this section intends to address. The first is to answer how one would consider correcting for multiplicity, irrespective of the *type* of correction. The meta-analyst must choose to correct across all the tests of statistical significance or to group the tests in some meaningful way. The solution to this problem is to create a diagnostic and decision tool to address when and at what stage the meta-analyst should consider correcting for multiplicity. The second problem, what type of correction one should

utilize, is addressed by using a variety of corrections available in the literature. Although this solution remains theoretically benign, the ease of computation and practical presentation outweigh the admitted weaknesses. The curious reader has many excellent sources for the computational and theoretical proofs of these corrections (see Chapter 2).

Timeline of Statistical Significance Testing

It is not clear when, how, or under what scenario the meta-analyst should correct for multiplicity (Cafri et al., 2010). Much like primary research, this is largely a philosophical question with no apparent answer: Should all tests of statistical significance be corrected simultaneously (i.e., experiment-wise correction) or only within certain groups of tests (i.e., family-wise correction)? There are three potential scenarios possible when considering a correction methodology (Table 17).

Table 17. Possible Correction Methodologies

Scenario	Challenges	Example
Across all tests within study	High rate of Type 2 error	A
Only "within-synthesis"	Does not consider experiment-wise error	B & C
Family-wise testing	More complicated, no clear methodology	Tables 22-25

The first possible scenario corrects across all tests within a given review (Table 18). Given "x" number of statistical significance tests, one could correct across all tests. For ease of use and interpretation, assume the use of the Bonferroni correction. Given 100 tests of statistical significance, the meta-analyst would divide the nominal alpha by 100, producing a critical value of .0005 (.05/100). Even given less conservative measures of correction than the Bonferroni correction, for instance the Holm procedure, still suffers from large a potentially large number of Type 2 errors in the process. Given the concern

for Type 2 errors and the inherent large numbers of statistical significance testing, it seems reasonable to assume experiment-wise correction is unacceptable.

Table 18. Example A: Combining All Significance Tests across the Review

Type of Test	Number of Tests	Percentage of Total
A1. Test of Overall Average	10	10%
A2. Test of Overall Homogeneity	10	10%
B1. Test of Subgroup Average	0	0%
B2. Test of Subgroup Homogeneity	0	0%
C1. Tests of Q-Between	20	20%
C2. Tests of Subgroup Average	0	0%
C3. Tests of Q-Within	60	60%
C4. Pairwise comparisons	0	0%
D1. Meta-regression Tests of Slopes	0	0%
Experiment-wise Total	100	100%
Bonferroni Correction	.05/100 = .0005	

Correcting across all statistical significance tests within one review disregards how many independent syntheses the analyst conducts. Another logical scenario, therefore, is to correct for multiplicity only "within" each independent synthesis. This solution is intuitively appealing, but to illustrate the problem inherent in this scenario, consider two examples. In the first example, a meta-analysis synthesizes 100 effect sizes across 10 independent syntheses (Table 19). Within each of these syntheses, the meta-analyst conducts a test of the overall mean, a test of the overall heterogeneity, and hypothesizes that two categorical variables moderate the variance of effect sizes, thus uses the *Q*-Between with six *Q*-Within tests of statistical significance (i.e., 3 levels per categorical variable). Each synthesis, as such, conducts a total of 10 tests of statistical significance. The experiment-wise number of statistical significance tests equals 100 (i.e., 10 syntheses * 10 tests per synthesis). Correcting "within" each synthesis, using the Bonferroni correction, however, assumes the nominal p-value cutoff is .005 (.05/10).

A second example decides to “lump” all effect sizes into one average effect rather than “split” the sample of effect sizes 10 different ways (Table 20). The meta-analyst, again, conducts a test of the overall mean, a test of the overall heterogeneity, two categorical variables using the *Q*-Between and *Q*-Within tests of statistical significance, each with 3 levels, for a total of 10 tests of statistical significance. Only one synthesis is conducted and therefore 10 tests of statistical significance are conducted. Using the Bonferroni correction, the nominal p-value cutoff is again .005 (.05/10).

Table 19. Example B: Combining Significance Tests within Multiple Syntheses

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
A1. Test of Average	1	1	1	1	1	1	1	1	1	1
A2. Test of Homogeneity	1	1	1	1	1	1	1	1	1	1
B1. Test of Subgroup Average	0	0	0	0	0	0	0	0	0	0
B2. Test of Subgroup Homogeneity	0	0	0	0	0	0	0	0	0	0
C1. Tests of Q-Between	2	2	2	2	2	2	2	2	2	2
C2. Tests of Subgroup Average	0	0	0	0	0	0	0	0	0	0
C3. Tests of Q-Within	6	6	6	6	6	6	6	6	6	6
C4. Pairwise comparisons	0	0	0	0	0	0	0	0	0	0
D1. Meta-regression Slopes Tests	0	0	0	0	0	0	0	0	0	0
Total Tests	10	10	10	10	10	10	10	10	10	10
Bonferroni Correction	.05/10 = .005	.05/10 = .005	.05/10 = .005	.05/10 = .005	.05/10 = .005	.05/10 = .005	.05/10 = .005	.05/10 = .005	.05/10 = .005	.05/10 = .005
Experiment-wise Total	10 Syntheses * 10 Tests per synthesis = 100									

Notes: S = Synthesis (e.g., S1 = Synthesis 1).

Should the second synthesis be treated in the same manner as the first? Under the scenario of "correcting within the synthesis", each synthesis utilizes the same procedure. However, as can be seen by the example above, the second synthesis has 10 times as many tests of statistical significance. Surely the second example must be treated differently than the first, but it is not apparent how to adjust the procedure given the parameters of correcting "within" each synthesis.

Table 20. Example C: Combining Significance Tests across One Synthesis

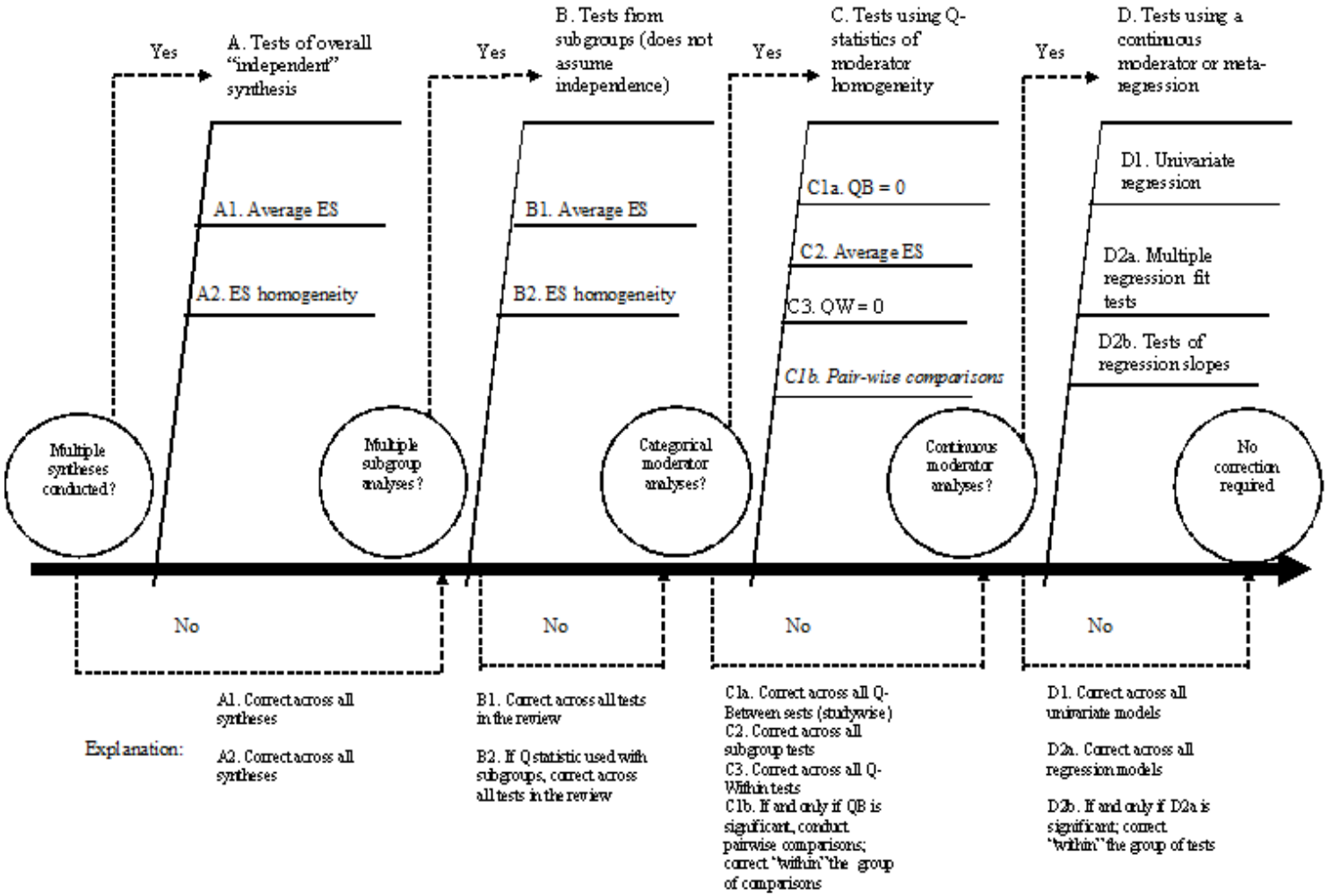
	One Synthesis	Percentage of Total
A1. Test of Overall Average	1	10%
A2. Test of Overall Homogeneity	1	10%
B1. Test of Subgroup Average	0	0%
B2. Test of Subgroup Homogeneity	0	0%
C1. Tests of Q-Between	2	20%
C2. Tests of Subgroup Average	0	0%
C3. Tests of Q-Within	6	60%
C4. Pairwise comparisons	0	0%
D1. Meta-regression Tests of Slopes	0	0%
Total Synthesis Tests	10	100%
Bonferroni Correction	.05/10 = .005	
Total Experiment-Wise Tests	1 Synthesis * 10 Tests = 10	

A third scenario, then, corrects only within a given type or set of tests (Figure 5), not experiment-wise or "within-synthesis"-wise. Given this scenario, tests of statistical significance must be corrected for multiplicity only within the group of tests (i.e., family) conducted concurrently. Each test is defined by the order in which the analyst conducts the test, hence the name the "timeline of statistical significance testing". The underlying conceit is that the meta-analyst uses each test of significance as a decision of whether to

conduct further tests of significance. For example, given a significant overall Q test of homogeneity, the meta-analyst is warranted to continue testing for moderators of the effect size. Accepting the null hypothesis of homogeneity, in contrast, the meta-analyst is advised to discontinue the search for significant moderators. Correcting within each type of test ensures such decisions are made systematically and efficiently. In a sense, the meta-analyst corrects row-wise across Table 19, but the "timeline" follows the logic of statistical significance testing in meta-analysis.

To assist in this procedure, the analyst answers a series of questions which then groups the statistical tests. The circles represent the major questions to be answered "yes" or "no". If an analyst answers "yes" to one of the questions, the tests addressed in each of the rows are grouped together. For example, the group "A" represents the set of independent syntheses and the overall tests of average effect and homogeneity. The analyst groups the entire set of tests of overall average effect, across the entire review. Similarly, but within a separate group, the analyst group the entire set of overall homogeneity tests. Group "B", then, represents subgroup analyses that may or may not be independent from one another. Group "C" refers to categorical moderator analyses that use the *Q*-Between/*Q*-Within approach. Group "D" relates to the procedures of meta-regression models. If the analyst only conducts one synthesis and no moderator analyses, then the answer to each question is "No" and no corrections for multiplicity are required.

Figure 5. Timeline of Statistical Significance Testing



The analyst starts by correcting "within" the group of tests that test the null hypothesis of average effect (i.e., H_0: $\theta = 0$, Group A1). For example, given Example B where the analyst "split" the effect sizes into ten independent syntheses, the meta-analyst, using the Bonferroni correction, uses the p-value cutoff of .005 (.05/10). Similarly, for the test of homogeneity (Group A2), the analyst simultaneously corrects for multiplicity across all tests of overall homogeneity. Using the Bonferroni correction, again, the cutoff p-value would be .005 (.05/10). Example C answers "No" to the first question, because only one independent syntheses is conducted, and therefore skips to the second question (Group B). No correction for multiplicity is required for either test.

The second question asks whether multiple subgroup analyses are conducted. To reiterate, this author refers to subgroup analyses where a meta-analyst groups effect sizes based on some category (i.e., population type) and uses a z-test of the overall effect size being different from zero. Subgroup analyses, under this definition, do not therefore include categorical moderator tests using the *Q*-Between tests of homogeneity methodology. These subgroup analyses are different from independent syntheses in one important way: The subgroups are not inherently independent from each other. Effect sizes may be represented in multiple subgroups. Hostetter's (2011) review provided an excellent example of subgroup analyses (pg. 307). The author grouped the effect sizes into 18 *dependent* subgroups, based on 8 "moderators". The first moderator was "speaking topic" and it was "subgrouped" into "abstract", "spatial", and "motor" (pg. 306). Each subgroup's effect sizes were synthesized to estimate an average effect size. The second moderator "gesture-speech redundancy" provided two subgroups:

"redundant" and "nonredundant". Again these subgroups' effect sizes were synthesized to estimate an average effect size. The reason these two moderators were dependent, however, was that the same effect sizes were used across the moderators. An independent synthesis would not utilize the same effect size across syntheses.

For examples B and C, however, the meta-analysts did not use subgroups to synthesize effect sizes. As such, each analyst would answer "No" to the second question and move to the third question. The next question asks if a Q-Between/Q-Within method of significance testing is utilized. Both example syntheses utilize this method of moderator analysis. Example B uses 2 Q-Between tests across 10 syntheses for a total of 20 tests of significance; therefore, the nominal cutoff, using the Bonferroni correction, is .0025 (.05./20). Example C conducts 2 Q-Between tests of significance, using the Bonferroni correction the nominal cutoff is .025 (.05/2). The same logic applies to the Q-Within tests of significance.

The advantage of using this format is apparent: Example C, where only one synthesis is conducted, is not punished as greatly relative to Example B, in terms of power, because it uses less experiment-wise tests of significance. Example B, given the large number of statistical tests, must incur some "punishment" for the increased use of significance testing.

Both examples did not use pairwise tests of significance, but it important to clarify the procedures here. Given pairwise comparisons, the meta-analyst corrects "within" the pairwise comparisons. The logic here follows that these comparisons are

inherently *independent* (i.e., the effect sizes are represented within only one of the subgroups being compared), given that they only test comparisons within the pairs.

Finally, the analyst answers the question for section D, the use of meta-regression. There are two types of regression models utilized in a meta-analysis. The first uses multiple univariate regression models, essentially testing each moderator independent of other moderators. This is a common practice and supported by some in the literature when the number of studies included in the review is small (i.e., $k < 10$; Higgins & Thompson, 2004). Under this scenario, then, the researcher includes each univariate model in the total number of significance tests. Given, for example, five univariate regression models, the Bonferroni corrected nominal cutoff is .005 (.05/5).

The second type is multiple regression. Here the meta-analyst simultaneously conducts tests of slopes most often using the weighted-least squares approach. The meta-analyst conducts two types of tests. The first tests the overall model-fit, a Q test. The reviewer corrects across all model-fit statistics. Given significant fit statistics for a given model, the analyst is then warranted to utilize the results of the regression model. Each of these slope tests must also be corrected for multiplicity, yet each are grouped by model.

Using Statistical Corrections in Meta-Analysis

As explained earlier, very few suggestions have been provided for the meta-analyst to consider with regard to the type of statistical correction to utilize. Hedges and Olkin (1985) suggested the Bonferroni corrections. An alternative to the Bonferroni correction, Holm's procedure, has been proposed and thus is relevant to these analyses. Scheffe's test is not available for direct corrections on p-values and thus Sidak's

correction was utilized as a close alternative. Finally, Laird et al (2010) advocated for the use of the FDR stepwise procedures. There is also an implicit suggestion not to correct for multiplicity whatsoever.

The following sections details 4 examples from the reviewed database. These samples of studies were derived first by coding whether all of the significance tests included in the review scribed the exact p-value. A surprisingly few number of studies (20.8%) actually provided the exact p-values. The examples listed below were chosen to illustrate different ways meta-analyses utilized statistical significance testing and how the timeline, in conjunction with the multiplicity corrections listed above, could serve to produce conclusions with greater statistical conclusion validity.

Abrami et al. (2008) Example. The first example, by Abrami et al. (2008), sought to synthesize intervention effect sizes that attempt to increase level of critical thinking in high school students. The review synthesized 161 effect sizes across 117 studies. The results indicated an overall positive impact of the intervention programs (g = .34) although the effect sizes were highly heterogeneous (Q = 1,767.86, p < .001).

The authors utilized only one independent synthesis, electing to "lump" all effect sizes into one group prior to synthesis (Table 21). This failed to abate the use of statistical significance tests, however, as 63 individual tests of significance were conducted (not counting redundant Q-Total or Q-Within tests). Of the 63 individual tests of significance, 57 reported p-values less than .00001. In fact, only 2 of the 63 (3.17%) tests of statistical significance had p-values above .05. Nevertheless, the example is still useful to detail the use of the timeline and the suggested corrections for meta-analysis.

Table 21. Abrami et al. (2008) Example Using the "Timeline"

Group of Tests	Number of Tests	Significant Tests	After Bonferroni	After Holm	After Sidak	After FDR	Text Changes
A1	1	1	-	-	-	-	0
A2	1	1	-	-	-	-	0
B1	27	26	26	26	26	26	0
B2	-	-	-	-	-	-	-
C1a	7	6	5	6	5	6	0-1
C2	-	-	-	-	-	-	-
C3	27	27	27	27	27	27	0
C1b	-	-	-	-	-	-	-
D1	-	-	-	-	-	-	-
D2a	-	-	-	-	-	-	-
D2b	-	-	-	-	-	-	-
Total	**63**	**62**	**59***	**60***	**59***	**60***	**0-1**

Notes: *Indicated that not all significance tests counted toward total because of lack of correction.

Because the review lumped all effect sizes into one group, only one test of statistical significance was utilized in group A1. Similarly, only one overall test of homogeneity was reported for group A2. The review then grouped the studies according to non-independent subgroups. For example, the authors organized the results by the type of research design, "Preexperimental *[sic]*", "Quasiexperimental *[sic]*", and "True experimental" and conducted significance tests of the average effect size for each subgroup (pg. 1115). As such, 27 statistical significance tests were conducted and 26 were reported as significant by the review authors. As the results indicated in Table 22, none of the null hypotheses were accepted even after correcting for multiplicity. This is not because the correction did not work or is unable to handle 27 tests, rather the p-values were remarkably small. The authors, it should also be mentioned, did not utilize a multiplicity correction.

The authors then conducted *Q*-Between/*Q*-Within tests of statistical significance. The review authors utilized 7 moderators, testing the homogeneity of effect sizes within each of the 7 moderators. As such, 7 *Q*-Between tests of significance were conducted. Of the 7 tests, the review authors concluded that 6 were statistically significant, again, without using a multiplicity correction. The results of the multiplicity corrections revealed that, using the most conservative correction (Bonferroni), at least 1 of the 6 should not be considered statistically significant. The FDR and Holm's procedures, however, indicated that all 6 should be considered statistically significant.

Following the timeline, the review authors also conducted twenty-seven tests of Q-Within. Similar to the subgroup analyses, none of the 27 p-values were considered non-significant, even after using the most conservative corrections. This was again attributed to the fact that all of the p-values were less than .0001. As such, a total of 63 tests of statistical significance were conducted. Of the 63, the review authors reported that 62 were statistically significant. After multiplicity corrections, only 1 of 62 would possibly be considered non-significant.

Dominguez et al. (2009) Example. The previous example lumped all effect sizes into one heterogeneous group before estimating subgroup and moderator effects. Dominguez and colleagues (2009) offered a different approach. The authors synthesized bivariate correlations of psychopathology and nuerocognition, observing the relationship across nine domains and four predictors. The review included 58 studies.

A total of 35 independent syntheses (one synthesis had only one effect size) were calculated. For each, an overall test of average effect was calculated. The authors

provided an I^2 statistic and therefore did not utilize a significance test of overall homogeneity. However, the authors conducted univariate meta-regression tests for three moderators (percent gender, average age, and chronicity) across all 35 syntheses. In sum, the authors conducted 140 tests of statistical significance. The authors did not use a multiplicity correction.

Table 22. Dominguez et al. (2009) Example Using the "Timeline"

Groups of Tests	Number of Tests	Significant Tests	After Bonferroni	After Holm	After Sidak	After FDR	Text Changes
A1	35	14	9	12	9	13	1-5
A2	-	-	-	-	-	-	-
B1	-	-	-	-	-	-	-
B2	-	-	-	-	-	-	-
C1a	-	-	-	-	-	-	-
C2	-	-	-	-	-	-	-
C3	-	-	-	-	-	-	-
C1b	-	-	-	-	-	-	-
D1	105	3	0	0	0	0	3
D2a	-	-	-	-	-	-	-
D2b	-	-	-	-	-	-	-
Total	**140**	**17**	**9**	**12**	**9**	**13**	**4-7**

To use the timeline, first the tests of average effect were grouped (Table 22). Using the traditional unadjusted cutoff of .05, 14 of the 35 tests of average effect were significant. The results of the Bonferroni correction reduced the number significant to 9. The Holm or FDR procedures have greater power relative to Bonferroni and therefore only reduce the number significant to 12 or 13, respectively. Given these results, between 1-5 changes to the authors conclusions would be required.

Because the authors utilized univariate regression, all meta-regression tests were grouped together. As such, 105 tests of statistical significance were utilized. Of the 105 tests, only 3 constituted statistical significance. Correcting for multiplicity using the

Bonferroni correction rendered the 3 tests non-significant (cutoff = .000005). In fact, none of the three tests would be statistically significant under any of the multiplicity corrections.

It should be noted that the authors discussed the 3 significant meta-regression slope tests. In each case, the authors stated that there was a high probability of Type 1 error and that the results were most likely spurious. The authors therefore disregarded the findings and did not recommend further postulation.

Archer (2000) Example. The meta-analytic literature in psychology produced many gender differences reviews. Archer's (2000) meta-analysis provided one such example, synthesizing gender difference effect sizes on aggressive behaviors in heterosexual couples. The authors collected 82 studies and conducted five independent syntheses; each synthesis rendered a unique set of effect sizes (k range = 14 – 82). The effect sizes were split by the outcome (self, partner, injury, and medical treatment) in addition to a "composite" variable that summarized each of the four outcomes within a primary study prior to synthesis.

For each independent synthesis, a test of the overall average effect size and overall homogeneity was calculated. The authors also conducted a series of one-way Q-Between tests for each of the five syntheses. The composite, self, and injury syntheses included a test of Q-Between for nine moderators. Each of the nine moderator tests included subgroup tests of the average effect size and Q-Within tests of significance. For the other two syntheses, the authors only conducted five Q-Between moderator tests of significance, but again conducted tests of subgroup average effect and Q-Within tests.

Finally, the authors conducted univariate and multiple regression models. Unfortunately, the authors failed to report exact p-values and therefore these models were not included in the final analysis. As such, Archer (2000) reported exact p-values for 181 tests of statistical significance.

Table 23. Archer (2000) Example Using the "Timeline"

Groups of Tests	Number of Tests	Significant Tests	After Bonferroni	After Holm	After Sidak	After FDR	Text Changes
A1	5	5	4	5	4	5	0-1
A2	5	5	5	5	5	5	0
B1	-	-	-	-	-	-	-
B2	-	-	-	-	-	-	-
C1a	35	22	14	15	15	20	2-8
C2	87	70	50	53	50	63	7-20
C3	47	27	24	24	24	25	2-3
C1b	-	-	-	-	-	-	-
D1	-	-	-	-	-	-	-
D2a	-	-	-	-	-	-	-
D2b	-	-	-	-	-	-	-
Total	**181**	**130**	**97**	**102**	**98**	**118**	**11-32**

Following the timeline, the first question asks if multiple syntheses were conducted. The authors conducted five independent syntheses and as such, these tests were grouped together (Table 23). The review author's results indicated that all five of the overall average effect sizes were statistically significant. The results of the multiplicity corrections indicated, however, that using the Bonferroni or Sidak correction would render one of the five tests non-significant. The second part of the first question (A2) failed to result in different conclusions.

The review author did not use subgroup analyses and therefore answers "no" to the second question. As such, the next question in the timeline asks about the use of Q-

Between tests for moderator examination. Thirty-five such tests of statistical significance were conducted across all 5 syntheses. Of the 35, 22 were reported as statistically significant in the published review. The results of the multiplicity corrections indicated that, at the least, two of those 22 tests would not have been significant. Under the strictest correction, Bonferroni, only 14 of the 22 tests would remain statistically significant. Similar results were found for the subgroup tests of average effect. Eighty-seven tests of statistical significance were conducted and 70 were reported as statistically significant. Following the correction procedure, at least 7 of the 70 would not be considered statistically significant under the least conservative model (FDR). Using the Bonferroni correction would result in only 50 of the 70 tests remaining statistically significant. The final set of tests, Q-Within, revealed only a few differences between the unadjusted and adjusted p-values. The FDR procedure revealed only two differences while the Bonferroni revealed three different results.

Hostetter (2011) example. A final example by Hostetter (2011) provided exact p-values for a regression model. The review authors synthesized studies that tested whether including hand gestures with speech increased a participant's understanding of a message. The meta-analysis collected 63 studies, each providing one effect size. Only one synthesis was conducted but multiple subgroup analyses were tested in addition to the regression model.

The review authors began by testing the overall average effect and overall homogeneity. Because these procedures occurred only once, the answer to the first question on the timeline was "no" (Table 24). The second question asks about subgroup

analyses. The review author conducted multiple tests of the subgroup's average effect size. A total of 18 tests of statistical significance were conducted and all 18 were statistically significant at a p-value less than .05. Using the Bonferroni or Sidak corrections, however, rendered only 16 of the 18 tests statistically significant. The more powerful and less conservative tests, Holm and FDR, on the other hand, failed to accept any null hypotheses.

Table 24. Hostetter (2011) Example Using the "Timeline"

Groups of Tests	Number of Tests	Significant Tests	After Bonferroni	After Holm	After Sidak	After FDR	Text Changes
A1	1	1	-	-	-	-	-
A2	1	1	-	-	-	-	-
B1	18	18	16	18	16	18	0-2
B2	-	-	-	-	-	-	-
C1a	-	-	-	-	-	-	-
C2	-	-	-	-	-	-	-
C3	-	-	-	-	-	-	-
C1b	-	-	-	-	-	-	-
D1	-	-	-	-	-	-	-
D2a	1	1	-	-	-	-	-
D2b	10	3	2	2	2	2	1
Total	**31**	**28**	**18***	**20***	**18***	**20***	**1-3**

Notes: *Indicated that not all significance tests counted toward total because of lack of correction.

The third question on the timeline asks about Q-Between tests of statistical significance. Hostetter (2011) did not conduct any such tests, and therefore answers "no". The final question asks about meta-regression. A single multiple regression model was conducted, therefore D1 and D2 do not apply. The last portion of the section D corrects for multiple slope tests within a single model. There were 10 tests of statistical significance, three of which were reported as statistically significant. All four of the

correction techniques, however, indicated that one of the statistically significant findings could be a product of Type 1 error.

A total of 31 tests of statistical significance were conducted by Hostetter (2011). Of them, 28 were reported statistically significant. Using the multiplicity corrections revealed that 1-3 of the significant results would not be considered statistically significant.

Summary

The results of this phase of the project revealed a high number of statistical significance tests conducted per review. Moreover, the results confirmed that a guide to grouping the statistical tests is required given the disparity and rate of test usage. Using both the timeline and a statistical correction would render a portion of the previously assumed significant tests non-significant. Taken together, the four studies would need to modify an average of 3.33 conclusions ($\sigma = 4.09$). Although this number is not overwhelming, it does reaffirm that the presence of false conclusions is possible. The previous examples illustrated that steps can be taken to reduce the risk of false conclusions.

CHAPTER FIVE

DISCUSSION

Overview

The use of statistical significance testing in meta-analysis, given the probability of Type 1 error, is an issue worth studying. Indeed, the call for greater observation and policy regarding the null hypothesis significance test use in meta-analysis has repeatedly been made (Bender et al., 2008; Cafri et al., 2010). Given the pervasiveness of meta-analysis, and especially the use of its results, all issues that relate to the validity of its findings are of paramount importance.

This project, therefore, sought to answer the calls to investigate multiplicity in meta-analysis. The primary goal and purpose was to quantify statistical significance testing across education and psychology meta-analyses, specifically within PB and RER, the top review journals in each respective field. After a systematic review of the titles and abstracts of each citation from 1986-2011, this author randomly selected 130 articles for inclusion. Each of the articles were coded for the use of statistical significance testing, number of independent syntheses, multiplicity corrections, and a host of other variables.

The results of the first phase of the project revealed several findings of interest. Across both journals and all years, the average meta-analysis utilized 70.82 (s = 94.20) tests of statistical significance. Greater use of null hypothesis testing was found for PB

relative to RER. In addition, the use of statistical testing has slowly increased over the past 25 years. No one type of statistical test (e.g., Q-Between) contributed more often to the overall sum as there remain many ways to test for moderation.

The results indicated that the reason for an increase in statistical significance testing was due to the number of synthesis splits and, in turn, the number of independent syntheses. Indeed, one important aspect of this project was to code how meta-analysts reasoned synthesis splitting. Bender et al. (2008) hypothesized that there were primarily four reasons for synthesis splitting: Outcomes, effect sizes, treatment groups, and time points. The results of this study revealed that, in fact, up to 10 reasons exist independently in the literature. These reasons included the predictor or subscale, comparison group, participant type, data source, statistical reason, and a variety of other reasons. A majority of the time, outcome differences were reasoned by the review authors as the need for synthesis splitting. The predictor or subscale, as well as comparison group and participant type, were also reported often as reasons. Effect size, treatment group, and time point, surprisingly, were rarely rationalized as reasons to split effect sizes into independent syntheses.

The final quantitative analysis for the full sample constituted a multiple regression model. The model investigated how contextual as well as multiplicity characteristics interacted to impact the total number of statistical significance tests. The results of the model revealed that the date of publication, number of studies included in the review, and number of independent syntheses all had a significant, positive relationship with the dependent variable.

To understand how tests of statistical significance were utilized and if an ad hoc multiplicity correction could be applied, this author purposively selected 4 reviews for further analysis. Prior to testing the applications of a multiplicity correction, however, this author devised a correction decision tree that grouped the tests by their respective type. Instead of correcting across all tests of statistical significance in a single review or within one synthesis within one review, the new decision tree grouped like-usage tests. This provided correction again Type 1 errors within the specific families. Because significance testing follows a timeline pattern (i.e., a significant test "allows" one to conduct more tests), this new procedure guards against subsequent tests of statistical significance.

Each of the tests of statistical significance, along with accompanying p-values, were coded. For each review, this author conducted multiplicity corrections for each group of tests. The results revealed differences between the reported conclusions and the conclusion after the multiplicity corrections. Conclusions would differ based on the number type of correction used; however, it was clear that some of the conclusions posited by the review authors were false.

Implications for Meta-Analysis

The results of this project revealed that meta-analysis methodologists must be aware of the increased reliance and usage of statistical significance testing. The remarkable ramification of this conclusion, ironically, is that meta-analysis is often lauded as removed from the discussion of statistical significance testing. Indeed, Schmidt (1996) believed that meta-analysis would render the use of statistical significance testing

null. The use of meta-analysis, Schmidt said, "reveals more clearly than ever before the extent to which reliance on significance testing has retarded the growth of cumulative knowledge in psychology" (pg. 116). Paradoxically, the increased use of meta-analysis may have inadvertently increased the use of statistical significance testing.

Given what previous research has indicated about Type 1 errors in primary research, meta-analysts must consider the possibilities and ramification of similar errors. There is little doubt that the false rejection of the null hypothesis occurs within these reviews. Although the rate at which one makes a false judgment may be slower relative to primary research due to the increased precision and power of the tests of statistical significance, it is increasingly more difficult to understand the false from the true. The use of meta-analyses' conclusion to inform policy and practice, across not only the disciplines of education and psychology but many other disciplines as well, make the validity of those conclusions of paramount importance. We must consider the ramifications of concluding wrongly and guard against this possibility, even if it means increased likelihood of occasionally accepting a false null hypothesis (i.e., Type 2 errors).

One simple way to guard against spurious findings, without the use of ad hoc or post hoc procedures, is to demand explicitly clear protocols. We often ask reviewers to delineate unambiguous research questions and logic models, but fail to require a distinction of the construct of interest and how it will be measured and synthesized. Moreover, meta-analyst should define decisions to lump or split effect sizes in the protocol as well. A clear definition of the parameters of the construct, including how primary studies measure it and with what types of people in specific settings, is a

minimum. These must be made clear prior to the search and retrieval phase because inclusion/exclusion decisions are often made based on these decisions. Repeatedly reviewing the protocol, explicitly defined, has the potential to guard against unexpected scenarios.

The other, more complicated and data-driven response, is to utilize multiplicity corrections along with the newly derived "Timeline". These measures will guard against the almost inevitable Type 1 error as well as provide greater robustness of conclusions. The means to correct against spurious findings, moreover, have never been more accessible. The methods utilized for this review required no more than the computational intensity required to perform the actual meta-analyses. A review that conducts multiple independent syntheses, and then conducts multiple moderator analyses, should consider utilizing multiplicity corrections to ensure robustness of conclusions. Given the concern for Type 2 errors, furthermore, this author would recommend using one of the newer multiplicity corrections (i.e., the FDR approach). The classical correction techniques, for instance the Bonferroni correction, are likely to result in a high number of Type 2 errors given their restrictive alpha levels. The results of the correction tests reiterated the restrictive nature of classical controls.

From a reporting standpoint, to understand better how the review split (or lumped) effect sizes, a standardized tool should be utilized. The standard PRISMA tool and MECIR standards do not go far enough to explicate how review authors present the findings. This is quickly evident from a cursory review of either PB or RER. It is often difficult to determine how the reviewers 1) derived the inclusion/exclusion criteria, 2)

determined whether to lump or split effect sizes. True, most reviews provide an inclusion/exclusion section and explicate the broad decision. However, rarely are examples provided to illustrate why one type of effect size was included while another left excluded. Further, the final lumped or split effect size product is evident, but the decisions that lead to the conclusion often lack for detail.

Table 25. Exemplary Table from Miller et al.'s (1991) Review

	Prospective		Cross-sectional		
			High-risk	Health Population	
	High risk	Healthy Population	Angiography	Case Control	Population
Structured Interview	4	4	13	5	1
Activity Survey	4	7	9	7	7
Self-report Measures	0	5	8	4	3

Notes: Recreated from Miller et al. (1991); Number in cells represent number of studies for each independent synthesis.

Miller et al.'s (1991) synthesis, detailed previously, provided an excellent example of how review authors may illustrate how the sample of effect sizes was split (Table 25). Instead of presenting each effect size in one descending column, the review authors detailed how each effect size was grouped according to the split levels. For example, the review authors split the effect sizes first by the type of time point, either prospective or cross-sectional. The major headings at the top of the chart provide the reader with this information. From there, the review authors split the sample again by the type of population, illustrated at the next "level" of the chart. For the cross-sectional designs, the "healthy population" is subdivided further, as illustrated by the third level. Finally, the type of measure further subdivides each of the five columns, subdivided by time, population, and type of study. Each cell, therefore, represented one of the fourteen independent syntheses.

Meta-analysis methodologists should also advocate for a standardized means to test moderators. At present, there are five ways to test for moderator effects: 1) Q-Between/Q-Within, 2) Subgroup analyses, 3) Pair-wise comparisons, 4) Univariate meta-regression, 5) Multiple meta-regression. From the standpoint of decreasing multiplicity in meta-analysis, multiple regression is the logical choice. The model decreases testing because it simultaneously conducts moderator tests, therefore reducing the rate of simple one-way ANOVA tests of significance. Moreover, the model inherently controls for all other predictors in the model, and therefore provides a more precise result. Two-way interactions, under-utilized in most meta-analyses, are also simple to devise and test in a multiple regression model. Finally, thanks to high-powered and efficient computer software, multiple predictor meta-regression models no longer put a strain on the analytical process.

Limitations

A number of limitations about this project should be stated. The nature of this study was observational and therefore causal statements should not be made. These observations, in addition, are subject to human error and subjectivity. Furthermore, this project did not devise new methods of adjusting p-values or critical values. This project, instead, simply quantified the problem to inform future research.

Another limitation is the possibility of selection bias. It is feasible that PB and RER attract meta-analyses where the literature is developed and plentiful, therefore leading review authors to assume it is reasonable to conduct a high rate of significance testing. This is partially confirmed by the fact that reviews that included more than the

average number of studies tended to conduct more significance tests. Similarly, review authors might choose to publish their findings in PB or RER because they included a large number of studies. Given these concerns, one should limit generalizability hypotheses.

One critical limitation to consider is whether Type 1 errors are a problem in meta-analysis at all. At least with regard to the overall average effect size, the goal is to collect and synthesize the population of effect sizes across every available resource. As such, it is not clear whether the distribution of effect sizes is subject to traditional frequentist logic inherent in primary study inference testing. What is not in question, however, is the Type 1 errors associated with conducting moderator analyses. The conclusions drawn from multiple one-way ANOVAs (i.e., Q-Between tests) should be considered highly suspect under the condition that many such tests are conducted.

Conclusions

The conceit of this project was to investigate whether statistical significance testing is a problem worth addressing in meta-analysis. It seems clear now, given the high rate of null hypothesis significance testing coupled with the egregious lack of correction, Type 1 errors can and will impact the validity of meta-analytic results. Meta-analysis methodologists and practitioners simply cannot afford to ignore the problem while promoting the promise of the paradigm's results. We must take action to prevent any doubt about the results of meta-analysis.

We, as a meta-analytic community, simply cannot allow practitioners of meta-analysis to abuse the null hypothesis significance test. Without such awareness and

discussion, meta-analytic results will remain as ambiguous and confounded as often are primary research results. Given the preventability of this threat, henceforth, meta-analysts must consider the issue of meta-analysis multiplicity paramount.

APPENDIX A

SCREENING TOOL FOR REVIEW OF TITLES AND ABSTRACTS

Question	Code	Explanation (if needed)
Title		
1. Does the title of the study indicate that the study focuses on a psychological or educational topic?		
2. Does the title of the study indicate that the study is a quantitative synthesis?		
Abstract		
1. Does the abstract discuss a broad psychological or educational topic?		
2. Does the abstract report the results of a quantitative synthesis?		
3. Are moderator or meta-regression results reported in the abstract?		

APPENDIX B

“UNSURE” SCREENING TOOL

Item	Answer (Yes/No)	Explanation
1. Is the review a quantitative synthesis?		
2. Does this review use a vote-counting technique?		
3. Are the results of the review presented within the article?		
Decision (Include/Discard)		

APPENDIX C

REVIEW CODING DOCUMENT

Question	Code
A. Basic Information	
1. First author	Last name, Initials
1b. Number of authors	
2. Date of Publication	
3. Title of article	
4. Publication Source	
5. Funded	1 = Yes, 2 = Not Mentioned
B. Study Information	
1. Topic of study	Broadly defined
2. Description of purpose	Author defined
3. Type of study	Bivariate, Diff., Efficacy, Effectiveness, Prediction
4. Cited synthesis authors	(i.e., Hedges & Olkin, 1985)
5. Update of previous review	1=Yes, 2=No
6. Graphical plot included	1=Yes, 2=No
7. Inclusion of grey literature	1=Yes, 2=No
8. Power analysis	1=Yes, 2=No
9. Primary study quality	1=Yes, 2=No
10. Cohen's ES classification	1=Yes, 2=No
11. Publication bias analysis	1=Yes, 2=No
12. Model specification	1=Yes, 2=No
C. Multiplicity Information	
1. Did the authors conduct multiple independent syntheses?	1=Yes, 2=No
2. List each synthesis and reason for split	
2a. Number of studies	
2b. Did they report z-test or CI?	List
2c. Conduct homogeneity test and if significant	
2c. Number of Q-Between tests and significant tests	
2d. Number of Q-Within tests and significant tests	
2e. Number of z-tests for moderators and number significant tests	
2f. Number of meta-regression tests and number of significant tests	
2g. Did the review conduct sensitivity analyses?	1=Yes, 2=No
D. Correcting for multiplicity	
1. Did the authors discuss the issue of multiplicity anywhere in the article?	1 = Yes, 2 = Not Mentioned
2. Did the authors adjust their alpha rates for any of the analyses?	1 = Yes, 2 = No
2b. If yes, what technique did the authors use?	List
3. What is the alpha level?	
4. Are all hypothesis test p-values reported?	1 = Yes, 2 = No

APPENDIX D

CODING TOOL FOR EXTRACTING EXACT P-VALUES

Question	Code
1. Type of test	Z, Q, other
2. Description of test	What was the author testing?
3. Test group	A1, B1, etc.
4. Test statistic (if reported)	Exact statistic
5. Degrees of freedom (if reported)	Exact statistic
6. P-value (if reported)	Exact statistic

REFERENCE LIST

Abrami, P. C., Bernard, R. M., Borokhovski, E., Wade, A., Surkes, M. A., Tamim, R., & Zhang, D. (2008). Instructional interventions affecting critical thinking skills and dispositions: A stage 1 meta-analysis. *Review of Educational Research, 78,* 1102-1134.

Agresti, A. (2009). *Categorical data analysis*. New York, NY: John Wiley & Sons.

Aickin, M., & Gensler, H. (1996). Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *American Journal of Public Health, 86,* 726-728.

Archer, J. (2000). Sex differences in aggression between heterosexual partners: a meta-analytic review. *Psychological Bulletin, 126,* 651-680.

Bender, R., Bunce, C., Clarke, M., Gates, S., Lange, S., Pace, N. L., & Thorlund, K. (2008). Attention should be given to multiplicity issues in systematic reviews. *Journal of Clinical Epidemiology, 61,* 857-865. doi: 10.1016/j.jclinepi.2008.03.004

Bender, R., & Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of Clinical Epidemiology, 54,* 343-349. doi: 10.1016/s0895-4356(00)00314-0

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal or the Royal Statistical Society, 57,* 289-300.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis, 1,* 385-402.

Borenstein, M. (2009). Effect sizes for continious data. In H. Cooper, Hedges, L.V., Valentine, J.C. (Ed.), *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage, Inc.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. New York, NY: John Wiley & Sons.

Bornstein, R. F. (1989). Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological Bulletin, 106*, 265.

Bowers, P. N., Kirby, J. R., & Deacon, S. H. (2010). The effects of morphological instruction on literacy skills a systematic review of the literature. *Review of Educational Research, 80*, 144-179.

Bus, A. G., Van Ijzendoorn, M. H., & Pellegrini, A. D. (1995). Joint book reading makes for success in learning to read: A meta-analysis on intergenerational transmission of literacy. *Review of Educational Research, 65*, 1-21.

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender differences in risk taking: A meta-analysis. *Psychological Bulletin, 125*, 367.

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research, 45*, 239-270. doi: 10.1080/00273171003680187

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Campbell, D. T., Stanley, J. C., & Gage, N. L. (1963). *Experimental and quasi-experimental designs for research*. New York, NY: Rand McNally.

Cohen, J. (1994). The earth is round ($p<.05$). *American Psychologist, 49*, 997-1001.

Connell, A. M., & Goodman, S. H. (2002). The association between psychopathology in fathers versus mothers and children's internalizing and externalizing behavior problems: a meta-analysis. *Psychological Bulletin, 128*, 746-773.

Connelly, B. S., & Ones, D. S. (2010). An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. *Psychological Bulletin, 136*, 1092-1122.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. New York, NY: Rand McNally

Cooper, H. (2010). *Research synthesis and meta-analysis* (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis*. New York, NY: Russell Sage Foundation.

Dominguez, M., Viechtbauer, W., Simons, C. J., van Os, J., & Krabbendam, L. (2009). Are psychotic psychopathology and neurocognition orthogonal? A systematic review of their associations. *Psychological Bulletin, 135*, 157-171.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of American Statistical Association, 56*, 52-64.

Eagly, A. H., & Johnson, B. T. (1990). Gender and leadership style: A meta-analysis. *Psychological Bulletin, 108*, 233-256.

Field, A. P. (2003). Can meta-analysis be trusted? *The Psychologist, 16*, 642-645.

Fisher, R. A. (1935). *The design of experiments*. New York, NY: Hafner Publising Company.

Goodyear-Smith, F. A., van Driel, M. L., Arroll, B., & Del Mar, C. (2012). Analysis of decisions made in meta-analyses of depression screening and the risk of confirmation bias: A case study. *BMC Medical Research Methodology, 12*, 76-88.

Grabe, S., & Hyde, J. S. (2006). Ethnicity and body dissatisfaction among women in the United States: a meta-analysis. *Psychological Bulletin, 132*, 622-640.

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests*. London, England: Taylor & Francis Group.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. London, England: Academic Press Inc.

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39-65.

Higgins, J., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine, 23*, 1663-1682.

Higgins, J. P. T., & Green, S. (2011). Cochrane handbook for systematic reviews of interventions. Retrieved from www.cochrane-handbook.org

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York, NY: John Wiley & Sons.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandanavian Journal of Statistics, 6*, 65-70.

Hostetter, A. B. (2011). When do gestures communicate? A meta-analysis. *Psychological Bulletin, 137*, 297-315.

Howell, D. C. (2006). *Statistical methods for psychology*. Independence, KY: Cengage Learning Tools.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. New York, NY: Sage Publications, Inc.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139-155.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods, 17*, 137-152.

Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparison wise Type I error control. *Psychological Methods, 4*, 58-69.

Keselman, H. J., Miller, C. W., & Holland, B. (2011). Many test of significance: New methods for controlling type 1 errors. *Psychological Methods, 16*, 420-431.

Kling, K. C., Hyde, J. S., Showers, C. J., & Buswell, B. N. (1999). Gender differences in self-esteem: A meta-analysis. *Psychological Bulletin, 125*, 470-500.

Laird, A. R., Fox, P. M., Price, C. J., Glahn, D. C., Uecker, A. M., Lancaster, J. L., . . . Fox, P. T. (2005). ALE meta-analysis: Controlling the false discovery rate and performing statistical contrasts. *Human Brain Mapping, 25*, 155-164. doi: 10.1002/hbm.20136

Lehmann, E. L., & Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics, 33*, 1138-1154.

Lipsey, M. W. (2007). Unjustified inferences about meta-analysis. *Journal of Experimental Criminology*, *3*, 271-279.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Thousand Oaks, CA: Sage Publications, Inc.

Littell, J. H., Corcoran, J., & Pillai, V. K. (2008). *Systematic reviews and meta-analysis*. Oxford, England: Oxford University Press.

Lohr, S. (1999). *Sampling: Design and analysis*. Independence, KY: Cengage Learning.

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences. In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*. New York, NY: The Russell Sage Foundation.

McCartney, K., Harris, M. J., & Bernieri, F. (1990). Growing up and growing apart: A developmental meta-analysis of twin studies. *Psychological Bulletin, 107*, 226-237.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference:Part I. *Biometrika, 20A*, 175-240.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

Oliver, M. B., & Hyde, J. S. (1993). Gender differences in sexuality: A meta-analysis. *Psychological Bulletin, 114*, 29-51.

Orwin, R. G., & Vevea, J. (2009). Evaluating Coding Decisions. In H. Cooper, Hedges, L.V., Valentine, J.C. (Ed.), *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage, Inc.

Pena, E. A., Habiger, J. D., & Wu, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *The Annals of Statistics, 39*, 556-583.

Pigott, T. D. (2012). *Advances in meta-analysis*. New York, NY: Springer.

Ross, J. A. (1988). Controlling variables: A meta-analysis of training studies. *Review of Educational Research, 58*, 405-437.

Scheffe, H. (1959). *The analysis of variance*. New York, NY: Wiley & Sons, Ltd.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115-129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of signifance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Sporer, S. L., Read, D., Penrod, S., & Cutler, B. (1995). Choosing, confidence, and accuracy: A meta-analysis of the confidence-accuracy relation in eyewitness identification studies. *Psychological Bulletin, 118*, 315-327.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research, 10*, 277-303.

Sutton, A. J., & Higgins, J. P. T. (2008). Recent developments in meta-analysis. *Statistics in Medicine, 27*, 625-650. doi: 10.1002/sim.2934

Swanson, H. L., & Hoskyn, M. (1998). Experimental intervention research on students with learning disabilities: A meta-analysis of treatment outcomes. *Review of Educational Research, 68*, 277-321.

Tendal, B., Nüesch, E., Higgins, J. P. T., Jüni, P., & Gøtzsche, P. C. (2011). Multiplicity of data in trial reports and the reliability of meta-analyses: Empirical study. *BMJ, 343*, 1-13. doi: 10.1136/bmj.d4829

Trikalinos, T. A., & Ioannidis, J. P. A. (2005). Assessing the evolution of effect sizes over time. In H. R. Rothstein, Sutton, A.J., & Borenstein, M. (Ed.), *Publication Bias in Meta-Analysis*. West Sussex, England: John Wiley & Sons, Inc.

Tukey, J. W. (1949). *Comparing individual means in the analysis of variance*. Unpublished Doctoral Dissertation. Princeton University. New Jersey.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Weir, M. C., Grimshaw, J. M., Mayhew, A., & Fergusson, D. (2012). Decisions about lumping vs. splitting of the scope of systematic reviews of complex interventions are not well justified: A case study in systematic reviews of health care professional reminders. *Journal of Clinical Epidemiology, 65*, 756-763.

Welton, N. J., Sutton, A. J., Cooper, N. J., Abrams, K. R., & Ades, A. E. (2012). *Evidence synthesis for decision making in healthcare*. West Sussex, England: John Wiley & Sons.

Williams, R. T. (2012). *Using robust standard errors to combine multiple estimates with meta-analysis*. Unpublished Doctoral Dissertation. Loyola University Chicago. Chicago, IL.

Wilson, D. B. (2009). Systematic Coding. In H. Cooper, Hedges, L.V., Valentine, J.C. (Ed.), *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage, Inc.

VITA

Josh Polanin was born and raised outside of Peoria, IL. He graduated from Metamora Township High School in 2002. Josh attended the University of Illinois Urbana-Champaign majoring in psychology and matriculated in December 2006. Josh started at Loyola University Chicago in August 2008. Throughout his tenure, Josh was funded in part by the National Science Foundation, Loyola's School of Education, and was awarded a dissertation fellowship from the Graduate School in May of 2013.

During his graduate career, Josh published papers in the *School Psychology Review, Journal of Youth and Adolescence, Social Work Research and Practice,* and *Research Synthesis Methods.* In addition, he presented work at various national conferences, including the American Educational Research Association, Society for Research on Educational Effectiveness, and Society for Research Synthesis Methods. He was also invited to teach courses on meta-analysis at the annual 2012 Campbell Collaboration Colloquium.

Josh held leadership roles during his tenure at Loyola as well. He served as the senior data analyst for a CDC-funded cluster-randomized trial of a school bullying prevention program. Starting in June of 2011, Josh has also held the position of managing editor for the Campbell Collaboration's Methods group.

Josh will begin an Institute of Education Sciences funded postdoctoral fellowship at Vanderbilt University's Peabody Research Institute starting in August of 2013.